Eine kurze Einführung

in Statistik, Maßtheorie

Stochastik und Integration

A. B. Zeidler

MARS-Gruppe

6. April 2021 (194 Seiten)

This material may be referenced, reproduced and distributed in whole or in part, in any medium physical or electronic, provided that (1) the terms of the *Open Publication License* are adhered to, that (2) this license or an incorporation of it by reference is displayed in the reproduction and that (3) the original author and a reference to the site of this document is clearly stated. I.e. any reproduction or distribution has to meet all the conditions set forth in the *Open Publication License*, v1.0 or later, given in section 18 (the latest version is presently available at www.opencontent.org/openpub). Distribution or derivation of the work for commercial purposes is prohibited, unless prior permission is obtained from the copyright holder.

Inhaltsverzeichnis

1 Symbole **3**

2 Vorbereitungen **4**

3 Kombinatorik **13**

4 Eindimensionale Statistik **18**

5 Zweidimensionale Statistik **26**

6 Multivariate Statistik **35**

7 Laplace Räume **41**

8 Spezielle Verteilungen **54**

9 Markov-Ketten **69**

10 Mengenringe **88**

11 Sigma-Algebren **105**

12 Unabhängigkeit **118**

13 Normalverteilung **125**

14 Konfidenzintervalle **138**

15 Hypothesentests **158**

16 Unabhängigkeitstests **172**

17 Integrationstheorie **182**

18 Open Publication License **193**

Kapitel 1

Symbole

Symbolverzeichnis:

$\mathbb{N}$	Menge der natürlichen Zahlen $\{ 0, 1, 2, \dots \}$
$a \dots b$	Menge der ganzen Zahlen $k \in \mathbb{Z}$ mit $a \leq k \leq b$
$\#M$	Zahl ($\in \mathbb{N} \cup \{ \infty \}$) der Elemente der Menge M
$\mathcal{P}(M)$	Potenzmenge $= \{ A \mid A \subseteq M \}$ der Menge M
Δ_n	$(n-1)$-dimensionaler Standardsimplex in $\mathbb{R}^n$
	also $\{ (t_1, \dots, t_n) \in \mathbb{R}^n \mid t_i \geq 0, t_1 + \dots + t_n = 1 \}$
$f(A)$	Bildmenge $\{ f(x) \mid x \in A \}$ von A unter $f : X \to Y$
$f^{-1}(B)$	Urbildmenge $\{ x \in X \mid f(x) \in B \}$ von B unter $f : X \to Y$
$\mathrm{im}(L)$	Bild $L(U)$ der linearen Abbildung $L : U \to V$
$\mathrm{kn}(L)$	Kern $\{ u \in U \mid L(u) = 0 \}$ der linearen Abbildung $L : U \to V$
$\mathrm{mat}_{m,n}(\mathbb{R})$	Menge der $(m \times n)$-Matrizen über $\mathbb{R}$
$\mathrm{mat}_n(\mathbb{R})$	Menge der quadratischen $(n \times n)$-Matrizen über $\mathbb{R}$
$\mathrm{row}_i(A)$	i-te Zeile ($\in \mathbb{R}^n$) der Matrix $A \in \mathrm{mat}_{m,n}(\mathbb{R})$
$\mathrm{col}_j(A)$	j-te Spalte ($\in \mathbb{R}^m$) der Matrix $A \in \mathrm{mat}_{m,n}(\mathbb{R})$
A^*	die transponierte Matrix von $A \in \mathrm{mat}_{m,n}(\mathbb{C})$
$A^\dagger$	die konjugiert-transponierte Matrix von $A \in \mathrm{mat}_{m,n}(\mathbb{C})$
$R[x]$	Polynomring über dem Ring R in der Variablen x

Danksagung:
Mein ganz besonderer Dank gilt Florian Timmermann für den guten Austausch und die zahlreichen Anmerkungen und Korrekturen, die sehr zum Gelingen des Skriptums beigetragen haben.

Kapitel 2

Vorbereitungen

Dieses Skriptum soll eine Einführung in die Wahrscheinlichkeitstheorie (Stochastik) bieten. Diese tritt sofort in Erscheinung, wenn man Statistik betreibt, aber nur Stichproben macht und keine Vollerhebung durchführt. Es stellen sich dann natürlich die Fragen: *Was soll ich denn jetzt für einen Durchschnitt nehmen?* und auch *Woher weiß ich jetzt eigentlich, dass mein Durchschnitt auch für die gilt, die ich gar nicht befragt habe?* Um ernsthaft Statistik betreiben zu können, braucht man also Stochastik. Und umgekehrt ist die Stochastik eine Fortführung der Statistik und damit nicht von dieser zu trennen.

Wir werden sehen, dass eine Formulierung der Stochastik auf der Ebene von Mengen am übersichtlichsten wird. Sobald die erhobenen Merkmale nicht mehr diskret liegen, sondern kontinuierlich sind, stellt sich sofort die Frage, wie man über diese summieren kann. Die Antwort darauf ist in der Maß- und Integrationstheorie zu finden. Um die Stochastik weiter zu führen, muss man eben diese entwickeln.

An der Schule genügt es zumeist, endliche Wahrscheinlichkeitsräume zu betrachten. Bei diesen gibt es natürlich keine Probleme mit Konvergenz und Kontinuität. Doch bereits beim Übergang von der Binomialverteilung zur Normalverteilung verlässt man die heile Welt der Laplace-Räume und begibt sich in den Sumpf der kontinuierlichen Wahrscheinlichkeitsräume. Doch wie der Statistiker Carter Alexander bemerkte: *Einem Schulmeister nur wenig oder sehr oberflächliches Wissen über Statistik zu vermitteln ist so, als drücke man einem Baby eine Rasierklinge in die Hand.*

Deswegen sollte man den Übergang von der Statistik zur Stochastik und weiter von den diskreten zu den kontinuierlichen Wahrscheinlichkeitsräumen gut studieren. Wir geben zwar oft keine Beweise, sondern verweisen lediglich auf die Literatur, gehen aber formal exakt vor. Und das bedeutet, dass wir an einigen Stellen in die technischen Details einsteigen müssen.

Die Kapitel sind nicht nach Schwierigkeit sortiert und müssen auch nicht von vorne nach hinten durchgelesen werden. Insbesondere die folgenden Vorbereitungen können für Nicht-Mathematiker etwas abschreckend sein. Es ist für einige Leser sicherlich einfacher den ersten Abschnitt zu überspringen und direkt mit der Statistik einzusteigen. Erst wenn Statistik und Laplace-Räume keine Probleme mehr bereiten, sollten die allgemeinen Wahrscheinlichkeitsräume angegangen werden. Und viele Vorbemerkungen werden erst für diese benötigt. Wenn wir zuvor schon Notationen verwenden, die aus den Vorbereitungen stammen, so weisen wir extra darauf hin.

Notation 2.1: Logische Vorbereitungen:
Die Mathematik gewinnt ihre Stärke aus der Verwendung einer formalen Sprache, mit der nicht nur Begründungen sondern eindeutige Beweise geführt werden können. Die Aussagenlogik (die Verwendung von *und*, *oder* und *nicht*) ist sicherlich jedem bekannt. Genauso wichtig ist aber die Verwendung von *Quantoren*: so schreiben wir $\forall\, x \in M : \varphi(x)$ wenn die Aussage $\varphi(x)$ für alle x aus der Menge M gilt und $\exists\, x \in M : \varphi(x)$ steht dafür, dass es (mindestens) ein $x \in M$ gibt, so dass $\varphi(x)$ erfüllt wird.
Bei Hypothesentests werden wir aber die Eintrittswahrscheinlichkeit der Verneinung der Hypothese betrachten. Wir wollen an dieser Stelle also kurz darauf eingehen, wie sich Quantoren unter Verneinung verhalten. Es ist klar, dass die verneinte Aussage *nicht für alle* $x \in M$ *gilt* $\varphi(x)$ sich offenbar umformulieren lässt, zu *es gibt (mindestens) ein* $x \in M$ *so dass nicht* $\varphi(x)$ *gilt.* Und genauso wird aus *es gibt kein* $x \in M$ *so dass* $\varphi(x)$ *gilt* wiederum *für alle* $x \in M$ *gilt nicht* $\varphi(x)$, da $\varphi(x)$ ja nur wahr oder falsch sein kann (tertium non datur). Formal bedeutet das die Äquivalenz der Aussagen

$$\begin{aligned} \neg\forall\, x \in M : \varphi(x) \quad &\Longleftrightarrow \quad \exists\, x \in M : \neg\varphi(x) \\ \neg\exists\, x \in M : \varphi(x) \quad &\Longleftrightarrow \quad \forall\, x \in M : \neg\varphi(x) \end{aligned}$$

Als Beispiel betrachten wir das Cauchykriterium einer Folge $(a_n) \subseteq \mathbb{R}$. Die Folge (a_n) ist eine Cauchyfolge, falls $\forall\, \varepsilon > 0\, \exists\, n_\varepsilon \in \mathbb{N}\, \forall\, m \geq n_\varepsilon\, \forall\, n \geq n_\varepsilon : |a_n - a_m| < \varepsilon$. Also ist (a_n) *keine* Cauchyfolge, falls $\exists\, \varepsilon > 0\, \forall\, n_\varepsilon \in \mathbb{N}\, \exists\, m \geq n_\varepsilon\, \exists\, n \geq n_\varepsilon : |a_n - a_m| \geq \varepsilon$. D.h. es gibt ein $\varepsilon > 0$ so dass es beliebig große Zahlenpaare (m, n) gibt, so dass a_m und a_n auseinander liegen: $|a_n - a_m| \geq \varepsilon$.

Die zweite Frage auf die wir kurz eingehen wollen, ist die der Vertauschbarkeit der Quantoren. Offenbar vertauschen die Allquantoren $\forall\, x \in M$ und $\forall\, y \in N$ untereinander, solange die Mengen M und N nicht von den Variablen x bzw. y abhängen. Dasselbe gilt für Existenzquantoren untereinander:

$$\begin{aligned} \forall\, x \in M\, \forall\, y \in N : \varphi(x, y) \quad &\Longleftrightarrow \quad \forall\, y \in N\, \forall\, x \in M : \varphi(x, y) \\ \exists\, x \in M\, \exists\, y \in N : \varphi(x, y) \quad &\Longleftrightarrow \quad \exists\, y \in N\, \exists\, x \in M : \varphi(x, y) \\ \exists\, x \in M\, \forall\, y \in N : \varphi(x, y) \quad &\Longrightarrow \quad \forall\, y \in N\, \exists\, x \in M : \varphi(x, y) \end{aligned}$$

Die letzte Formel besagt, dass man Allquantoren vor Existenzquantoren vorziehen darf (wenn es ein x gibt so dass für alle y gilt $\varphi(x, y)$, dann hat auch jedes y ein $x(y) = x$ so dass $\varphi(x, y)$ gilt). Die Umkehrung davon gilt aber nicht! Das ist z.B. der Unterschied zwischen gleichmäßiger und punktweiser Konvergenz von Funktionen $f_n : M \to \mathbb{R}$. Gleichmäßige Konvergenz $(f_n) \to f$ liegt vor, wenn $\forall\, \varepsilon > 0\, \exists\, n_\varepsilon \in \mathbb{N}\, \forall\, x \in M\, \forall\, n \geq n_\varepsilon : |f(x) - f_n(x)| < \varepsilon$. Bei der punktweisen Konvergenz $(f_n) \to f$ haben wir lediglich $\forall\, x \in M\, \forall\, \varepsilon > 0\, \exists\, n_\varepsilon \in \mathbb{N}\, \forall\, n \geq n_\varepsilon : |f(x) - f_n(x)| < \varepsilon$. Es ist bekannt, dass die punktweise Konvergenz aus der gleichmäßigen Konvergenz (durch Vorziehen von $\forall\, x \in M$) folgt. Speziell für stetige Funktionen $f_n : M \to \mathbb{R}$ und kompakte Intervalle $M \subseteq \mathbb{R}$ gilt auch die Umkehrung, im allgemeinen ist das aber falsch.

Notation 2.2: Algebraische Vorbereitungen:
Bekanntlich besteht die Menge $\mathbb{R}^n$ (für beliebiges $1 \leq n \in \mathbb{N}$) einfach aus n Kopien von $\mathbb{R}$ und deren Elemente werden als n-Tupel $x = (x_1, x_2, \ldots, x_n)$ geschrieben. Dabei ist jedes $x_k \in \mathbb{R}$ ($k \in 1 \ldots n$). Anders gesagt: $\mathbb{R}^n$ ist das kartesisches Produkt von n Kopien von $\mathbb{R}$. Wir schreiben die Elemente von $\mathbb{R}^n$ wahlweise als Spalten- oder Zeilenvektoren, je nachdem welche Darstellungsform günstiger ist - einen inhaltlichen Unterschied gibt es dabei nicht.

$$(x_1, x_2, \ldots, x_n) \quad = \quad \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$$

Ist $x = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$ ein n-Tupel, dann schreiben wir $[x]_i := x_i$ für die i-te Komponente von x. Es ist eine bekannte Tatsache, dass $\mathbb{R}^n$ zum $\mathbb{R}$-Vektorraum wird, unter der punktweisen Addition und Skalarmultiplikation. Sie wird sogar zur (kommutativen) $\mathbb{R}$-Algebra (d.h. einem Vektorraum mit gewohnter Multiplikation) unter der punktweisen Multiplikation. D.h. sind $x = (x_1, x_2, \ldots, x_n)$, $y = (y_1, y_2, \ldots, y_n) \in \mathbb{R}^n$ und ist $a \in \mathbb{R}$, dann setzen wir:

$$\begin{aligned} x + y &:= (x_1 + y_1, x_2 + y_2, \ldots, x_n + y_n) \\ xy &:= (x_1 y_1,\ x_2 y_2,\ \ldots,\ x_n y_n) \\ ax &:= (a x_1,\ a x_2,\ \ldots,\ a x_n) \end{aligned}$$

Während diese Konstruktionen absoluter Standard sind, fixieren wir auch eine Einbettung von $\mathbb{R}$ in $\mathbb{R}^n$, indem wir eine Zahl $a \in \mathbb{R}$ stets auch als Diagonalelement $a = (a, a, \ldots, a) \in \mathbb{R}^n$ auffassen:

$$\mathbb{R} \ \hookrightarrow \ \mathbb{R}^n \ : \ a \ \mapsto \ (a, a, \ldots, a)$$

Man bemerke, dass dies konsistent zur Konstruktion der $\mathbb{R}$-Algebra $\mathbb{R}^n$ ist, da $ax = (ax_1, ax_2, \ldots, ax_n)$ dasselbe Ergebnis liefert, egal, ob wir $a \in \mathbb{R}$ oder $a \in \mathbb{R}^n$ betrachten. Schließlich ist auch bekannt, dass auf $\mathbb{R}^n$ ein kanonisches Skalarprodukt existiert:

$$x \bullet y \quad := \quad x_1 y_1 + x_2 y_2 + \ldots + x_n y_n$$

Und als solches induziert das Skalarprodukt eine natürliche Norm und diese wiederum eine Metrik auf $\mathbb{R}^n$. Wir geben deren Konstruktionen der Vollständigkeit halber an:

$$\begin{aligned} \|x\| &:= \sqrt{x \bullet x} \ = \ \sqrt{x_1^2 + \ldots + x_n^2} \\ d(x, y) &:= \|x - y\| \end{aligned}$$

Notation 2.3: Topologische Vorbereitungen:
Auf dem Körper $\mathbb{R}$ existiert eine totale Ordnungsrelation $\leq$, die definiert wird, durch ihren positiven Kegel $a \geq 0 \iff \exists\, r \in \mathbb{R} : a = r^2$. Äquivalent dazu gibt es einen multiplikativen Betrag, der definiert wird, durch:

$$\begin{aligned} \operatorname{sgn}(a) &:= \begin{cases} 1 & \text{für } a > 0 \\ 0 & \text{für } a = 0 \\ -1 & \text{für } a < 0 \end{cases} \\ |a| &:= \begin{cases} a & \text{für } 0 \leq a \\ -a & \text{für } a \leq 0 \end{cases} \end{aligned}$$

Bekanntlich sind die reellen Zahlen vollständig und daher besitzt auch jede nach oben beschränkte Teilmenge von $\mathbb{R}$ ein Supremum. D.h. ist $X \subseteq \mathbb{R}$ eine nicht-leere ($X \neq \emptyset$) Teilmenge, mit: $\exists\, m \in \mathbb{R}$ so dass $\forall\, x \in X$ gilt $x \leq m$, dann gibt es ein (eindeutig bestimmtes) $s \in \mathbb{R}$ mit

(1) s ist eine obere Schranke von X: $\forall\, x \in X : x \leq s$

(2) s ist minimal unter diesen: $\forall\, m \in \mathbb{R}$ gilt $\forall\, x \in X : x \leq m \implies s \leq m$

Dieses $s \in \mathbb{R}$ wird **Supremum** von X genannt, geschrieben als $s := \sup(X)$. Analog besitzt jede nicht-leere, nach unten beschränkte Teilmenge $X \subseteq \mathbb{R}$ eine größte untere Schranke, die als **Infimum** bezeichnet wird. Man könnte die obigen Eigenschaften analog angeben oder kurzerhand $\inf(X) = -\sup(-X)$ verwenden. Wir wollen aber *jeder* Teilmenge $X \subseteq \mathbb{R}$ ein Supremum und Infimum zuordnen. Dazu ist es notwendig die reellen Zahlen um zwei Symbole zu erweitern, zu:

$$\mathbb{R} \cup \{\, \infty, -\infty \,\}$$

Nun erweitern wir die Ordungsrelation, Addition und Multiplikation (so gut es geht) auf diesen erweiterten Zahlenbereich. Man beachte, dass die erweiterte Ordnung $\leq$ weiterhin total ist, aber die algebraischen Operationen Definitionslücken lassen: $\mathbb{R} \cup \{\, \infty, -\infty \,\}$ ist kein Körper/Ring mehr

$$\begin{aligned} \forall\, a \in \mathbb{R} \cup \{\, -\infty \,\} &: a < \infty \\ \forall\, a \in \mathbb{R} \cup \{\, \infty \,\} &: -\infty < a \\ \forall\, a \in \mathbb{R} \cup \{\, \infty \,\} &: a + \infty := \infty \\ \forall\, a \in \mathbb{R} \cup \{\, -\infty \,\} &: a + (-\infty) := -\infty \\ \forall\, a \in \mathbb{R} \cup \{\, \pm\infty \,\} \setminus \{\, 0 \,\} &: a\infty := \operatorname{sgn}(a)\infty \\ \forall\, a \in \mathbb{R} \cup \{\, \pm\infty \,\} \setminus \{\, 0 \,\} &: a(-\infty) := -\operatorname{sgn}(a)\infty \end{aligned}$$

Dies erlaubt nun die Einführung von Supremum und Infimum für beliebige Teilmengen X der reellen Zahlen. Wir definieren:

$$\begin{aligned} \sup(X) &:= \begin{cases} -\infty & \text{für } X = \emptyset \\ \infty & \text{für } \nexists\, m \in \mathbb{R} : \forall\, x \in X : x \leq m \\ \sup(X) & \text{für } X \neq \emptyset \text{ und } \exists\, m \in \mathbb{R} : \forall\, x \in X : x \leq m \end{cases} \\ \inf(X) &:= \begin{cases} \infty & \text{für } X = \emptyset \\ -\infty & \text{für } \nexists\, m \in \mathbb{R} : \forall\, x \in X : m \leq x \\ \inf(X) & \text{für } X \neq \emptyset \text{ und } \exists\, m \in \mathbb{R} : \forall\, x \in X : m \leq x \end{cases} \end{aligned}$$

Wir erweitern nun die Ordnungsrelation $\leq$ von $\mathbb{R}$ auf $\mathbb{R}^n$. Man beachte, dass dabei aber nur eine partielle Ordnung entsteht (z.B. sind $(0,1)$ und $(1,0)$ unvergleichbar). Seien also allgemeiner $a = (a_1, a_2, \ldots, a_n)$ und $b = (b_1, b_2, \ldots, b_n) \in (\mathbb{R} \cup \{\pm\infty\})^n$, dann schreiben wir:

$$a \leq b \quad :\Longleftrightarrow \quad a_1 \leq b_1 \quad \text{und} \ \ldots \ \text{und} \quad a_n \leq b_n$$

$$a < b \quad :\Longleftrightarrow \quad a_1 < b_1 \quad \text{und} \ \ldots \ \text{und} \quad a_n < b_n$$

Vorsicht: diese Notation ist nützlich, aber $a \leq b$ ist hier *nicht* länger äquivalent, zu $a < b$ oder $a = b$. Beispielsweise ist $(0,0) \leq (0,1)$ aber es gilt weder $(0,0) < (0,1)$ noch $(0,0) = (0,1)$. Und damit können wir dann schließlich die Quader definieren:

$$\begin{aligned}
[a,b] &:= \{\, x \in (\mathbb{R} \cup \{\pm\infty\})^n \mid a \leq x \text{ und } x \leq b \,\} \\
[a,b[&:= \{\, x \in (\mathbb{R} \cup \{\pm\infty\})^n \mid a \leq x \text{ und } x < b \,\} \\
]a,b] &:= \{\, x \in (\mathbb{R} \cup \{\pm\infty\})^n \mid a < x \text{ und } x \leq b \,\} \\
]a,b[&:= \{\, x \in (\mathbb{R} \cup \{\pm\infty\})^n \mid a < x \text{ und } x < b \,\}
\end{aligned}$$

Definition 2.4: Folgen und Reihen:
Wir wollen kurz die Begriffe *Folge* und *Reihe* wiederholen: Eine **(reelle) Folge** a ist einfach eine Abbildung $a : \mathbb{N} \to \mathbb{R}$. D.h. jeder natürlichen Zahl n wird eine reelle Zahl $a_n := a(n)$ zugeordnet. In Anlehnung an n-Tupel schreibt man die Folge dann auch als $(a_n) := a$. Wir sagen die Folge (a_n) **konvergiert** gegen den *Grenzwert* $g \in \mathbb{R}$, falls

$$\forall\, \varepsilon > 0 : \exists\, n_\varepsilon \in \mathbb{N} \text{ so dass } \forall\, n \geq n_\varepsilon : |g - a_n| < \varepsilon$$

Dabei muss man nicht jeden Wert $\varepsilon > 0$ betrachten, sondern es genügt wiederum $\varepsilon = 1/m$ für beliebig große $m \in \mathbb{N}$. Die Vollständigkeit von $\mathbb{R}$ bedeutet gerade, dass eine Folge bereits dann konvergiert, wenn sie das *Cauchy-Kriterium* erfüllt (das den Grenzwert g nicht benötigt):

$$\forall\, \varepsilon > 0 : \exists\, n_\varepsilon \in \mathbb{N} \text{ so dass } \forall\, m, n \geq n_\varepsilon : |a_m - a_n| < \varepsilon$$

Wenn der Grenzwert g existiert, ist er eindeutig und wir schreiben $(a_n) \to g$. Die meisten Folgen divergieren jedoch, d.h. haben keinen Grenzwert.

Man kann auch Folgen in beliebigen metrischen Räumen (X, d) definieren, indem man den Abstand $|g - a_n|$ verallgemeinert, zu $d(g, a_n)$. Dies gilt insbesondere für $\mathbb{R}^n$. Aber Vorsicht: es gibt zahlreiche metrische Räume, in denen Folgen, die der Cauchy-Bedingung genügen, keinen Grenzwert haben müssen. Das bekannteste Beispiel ist $X = \mathbb{Q}$.

Ist $a = (a_n) \subseteq \mathbb{R}$ eine reelle Folge, dann kann man auch beginnen über die Folgeglieder zu summieren. In diesem Fall erhält man eine Folge (s_n) der Partialsummen

$$s_n \quad := \quad \sum_{i=0}^{n} a_i$$

Wir sagen die **Reihe** über (a_n) konvergiert, wenn die Folge (s_n) der Partialsummen konvergent ist. In diesem Fall nennen wir den Grenzwert der Folge der Partialsummen den Grenzwert der Reihe und schreiben:

$$\sum_{i=0}^{\infty} a_i \;:=\; g \quad \text{falls} \quad \left(\sum_{i=0}^{n} a_i\right) \to g$$

Das bekannteste Beispiel ist (q^n) für eine reelle Zahl $q \in \mathbb{R}$. Man rechnet leicht nach, dass in diesem Fall $s_n = (1 - q^{n+1})/(1 - q)$ ist. Für $|q| < 1$ konvergiert $(q^{n+1}) \to 0$ und damit $(s_n) \to 1/(1-q)$. Dies ist die *harmonische Reihe*

$$\sum_{i=0}^{\infty} q^i \;=\; \frac{1}{1-q} \quad \text{für} \quad |q| < 1$$

Definition 2.5: Summierbare Zahlenfamilien:
Wir wollen nun den Begriff der Reihen auf beliebige Indexmengen verallgemeinern. D.h. wir betrachten eine beliebige nicht-leere Menge $I \neq \emptyset$ und eine Zahlenfamilie (Funktion) $a : I \to \mathbb{R}$. Wie zuvor schreiben wir $a_i := a(i)$ und $(a_i) := a$. Ist I endlich, dann ist die Summe

$$\sum_{i \in I} a_i \;\in\; \mathbb{R}$$

wohldefiniert - da die Addition auf $\mathbb{R}$ kommutativ und assoziativ ist, kommt es nicht auf die Reihenfolge an, in der wir die Summe ausführen. Ist I unendlich, müssen wir die Menge der endlichen Teilmengen von I betrachten:

$$\Omega(I) \;:=\; \{\, E \subseteq I \mid \#E < \infty \,\}$$

Mit dieser Bezeichnung sind die folgenden Aussagen über $(a_i) \subseteq \mathbb{R}$ äquivalent (für einen Beweis und weitere Konsequenzen siehe [Barner, Fohr, Analysis I, Kapitel 5.5])

(a) Die Menge $\left\{ \sum_{i \in E} a_i \mid E \in \Omega(I) \right\} \subseteq \mathbb{R}$ ist beschränkt

(b) Die Menge $\left\{ \sum_{i \in E} |a_i| \mid I \in \Omega(I) \right\} \subseteq \mathbb{R}^+$ ist beschränkt

(c) Zu jedem $\varepsilon > 0$ gibt es eine endliche Teilmenge $E(\varepsilon) \in \Omega(I)$ so dass für alle anderen endlichen Teilmengen $F \in \Omega(I)$ gilt

$$E(\varepsilon) \cap F = \emptyset \quad \Longrightarrow \quad |\sum_{i \in F} a_i| < \varepsilon$$

(d) Es gibt ein $g \in \mathbb{R}$ so dass es zu jedem $\varepsilon > 0$ eine endliche Teilmenge $E(\varepsilon) \in \Omega(I)$ gibt, so dass für alle anderen $E \in \Omega(I)$ gilt

$$E(\varepsilon) \subseteq E \quad \Longrightarrow \quad |g - \sum_{i \in E} a_i| < \varepsilon$$

Ist eine (und damit alle) dieser Bedingungen erfüllt, dann nennen wir (a_i) **summierbar**. In diesem Fall ist der Grenzwert g in Aussage (d) eindeutig bestimmt. Für endliche I ist g natürlich die Summe der a_i und daher definieren wir allgemein:

$$\sum_{i \in I} a_i \quad := \quad g$$

Bemerkung: Wer glaubt, man könne auf diese Weise um die Integralrechnung herum kommen, der täuscht sich leider: Ist (a_i) summierbar, dann ist die Menge $\{\, i \in I \mid a_i \neq 0 \,\}$ der von 0 verschiedenen Einträge in (a_i) abzählbar. D.h. im Grunde hat man doch nur eine absolut konvergente Reihe vorliegen. Der Vorteil dieses Begriffes ist jedoch, dass er endliche Summen mit einschließt und man zuvor keine Abzählung fixieren muss.

Pseudozufallszahlen:
Für mathematische Experimente benötigt man oft große Mengen zufällig erzeugter Zahlen, hat aber als Hilfsmittel nur einen deterministisch arbeitenden Computer. Das Problem besteht also darin Zahlenfolgen zu erzeugen, die genauso zufällig wirken wie echte Zufallsergebnisse, sich aber berechnen lassen. Ein erstes Beispiel wäre die Zahlenfolge 3, 1, 4, 1, 5, 9... der Ziffern von π. Diese ist wunderbar regellos, gleichverteilt (das ist noch unbewiesen, wird aber stark vermutet) und lässt sich gut berechnen. Leider ist zu dieser Art Probleme noch sehr wenig bekannt. Deswegen wollen wir hier 3 Verfahren vorstellen, die verlässlich solche *Pseudozufallszahlen* liefern:

- Wir geben $n \in \mathbb{N}$ vor und betrachten die n-elementige Menge $Z := \mathbb{Z}_n = \{\, 0, 1, \ldots, n-1 \,\}$. Wählen wir nun zwei Parameter a und b und einen Startwert $z_0 \in Z$ dann definieren wir die Rekursionsvorschrift ($k \in \mathbb{N}$)

  $$z_{k+1} \quad := \quad (az_k + b) \mod n \quad \in \quad N$$

  Anders gesagt: $z_k = L^k(z_0)$ unter der Abbildung $L : \mathbb{Z}_n \to \mathbb{Z}_n$ definiert durch $L(x) := ax+b$. Offensichtlich ist die damit generierte Zahlenfolge $(z_1, z_2, z_3, \ldots) \subseteq \mathrm{im}(L) \subseteq \mathbb{Z}_n$ im Bild von L enthalten und damit endlich. Damit wiederholt sich diese Zahlenfolge (ab einer gewissen Stelle) sogar periodisch und ist daher (ab dieser Stelle) gleichverteilt. Leider trifft diese Folge nicht alle Elemente von N (selbst wenn a und n teilerfremd gewählt werden), so dass diese Methode noch ziemlich mangelhaft ist. Dennoch kann sie manchmal ganz gut funktionieren: Das folgende Beispiel ist für $n = 17$ mit $L(x) = 3x + 1$ berechnet worden. Als Startwert diente $z_0 = 2$, dann ergibt sich die Zahlenfolge

  $$(2, 7, 5, 16, 15, 12, 3, 10, 14, 9, 11, 0, 1, 4, 13, 6, 2, \ldots)$$

- *XOR-Shift:* Der folgende Algorithmus überzeugt durch seine einfache Implementierung und guten Ergebnisse. Die generierten Zufallsfolgen sind zwar wieder periodisch, doch liegt die Periodenlänge mit $2^{128} - 1 \approx 3.4 \cdot 10^{38}$ ziemlich hoch. Die generierten Zufallsvariablen liegen in $Z := 0 \ldots (2^{32} - 1)$ und verwenden die Binärdarstellung: Sind x_0 bis

$x_n \in \mathbb{Z}_2 = \{0, 1\}$ dann schreiben wir

$$(x_n \dots x_1 x_0)_b \;:=\; \sum_{i=0}^{n} x_i 2^i \;\in\; 0 \dots (2^{n+1} - 1)$$

Auf dieser Menge $\mathbb{Z}_2^n$ müssen wir nun ein paar klassische Rechenoperationen der Informatik einführen, den *Links-* und *Rechts-Shift* und die bitweisen *and* und *xor*-Verknüpfungen: Seien $0 \leq x, y < 2^{n+1}$ also $x = (x_n \dots x_1 x_0)_b$ und $y = (y_n \dots y_1 y_0)_b$ deren Binärdarstellungen, und sei $k \in \mathbb{N}$, dann

$$\begin{aligned}
\mathrm{shr}(x,k) &:= \mathrm{int}\left(\frac{x}{2^k}\right) = (0\,0 \dots 0 x_n \dots x_k)_b \\
\mathrm{shl}(x,k) &:= (2^k \cdot x) \bmod 2^{n+1} = (x_{n-k} \dots x_0\, 0\, 0 \dots 0)_b \\
x \;\mathrm{xor}\; y &:= ((x_n + y_n) \dots (x_0 + y_0)\,(x_0 + y_0))_b \\
x \;\mathrm{and}\; y &:= ((x_n \cdot y_n) \dots (x_1 \cdot y_1)\,(x_0 \cdot y_0))_b
\end{aligned}$$

Dabei sind $+$ und $\cdot$ die üblichen Operationen in $\mathbb{Z}_2$. Der folgende Algorithmus wurde http://www.jstatsoft.org/v08/i14/paper entnommen: er beginnt mit 4 vorgegebenen 32-bit Zahlen $z_k \in Z := 0 \dots (2^{32} - 1)$, wobei $k \in 0 \dots 3$, die nicht alle gleichzeitig 0 sein dürfen. In jedem Schritt $k \geq 4$ wird nun berechnet:

$$\begin{aligned}
t &:= z_{k-4} \;\mathrm{xor}\; \mathrm{shr}(z_{k-4}, 15) \\
z_k &:= (z_{k-1} \,\mathrm{xor}\, \mathrm{shl}(z_{k-1}, 21)) \;\mathrm{xor}\; (t \,\mathrm{xor}\, \mathrm{shl}(t, 4))
\end{aligned}$$

Dann ist $(z_0, z_1, z_2, z_3, z_4, z_5, \dots)$ die Folge der Zufallszahlen zum Keim (random seed) (z_0, z_1, z_2, z_3). Man kann diesen Algorithmus auch leicht mit nur 5 Speicherplätzen t, w, x, y und z programmieren, wobei z dann die jeweils nächste Zahl in der Folge der Zufallszahlen erhält:

$$\begin{aligned}
t &:= w \;\mathrm{xor}\; \mathrm{shr}(w, 15) \\
w &:= x \\
x &:= y \\
y &:= z \\
z &:= (z \,\mathrm{xor}\, \mathrm{shl}(z, 21)) \;\mathrm{xor}\; (t \,\mathrm{xor}\, \mathrm{shl}(t, 4))
\end{aligned}$$

- *Mersenne-Twister:* dieser Algorithmus ist zwar etwas komplizierter, aber dafür sind viele gute Dinge darüber bewiesen: Die generierten Zufallszahlen sind gleichverteilt, die Periodenlänge ist mit $2^{19937} - 1$ ziemlich groß und er besteht zahlreiche statistische Tests auf zufällige Verteilung. Er operiert ebenfalls auf der Menge $Z := 0 \dots (2^{32} - 1)$ der 32-bit Zahlen. Man beachte, dass diese als 8-stellige Hexadezimalzahlen dargestellt werden können, was wir im Algorithmus ausnutzen werden. Als Startwerte werden $n = 624$ echte Zufallszahlen y_1 bis $y_n \in Z$ gewählt. Diese stammen idealer Weise aus echten Zufallsprozessen - Wetterdaten, radioaktivem Zerfall oder anderen Quantenprozessen.

Die Binärdarstellung von y_k sei $(y_k^{31} \dots y_k^1 y_k^0)_b$. Aus diesen generiert man dann sukzessive (für $k > n$) die weiteren Zahlen:

$$\begin{aligned} h_k &:= y_{k-n} - (y_{k-n} \bmod 2^{31}) + (y_{k-n+1} \bmod 2^{31}) \\ &= (y_{k-n}^{31}\, y_{k-n+1}^{30} \cdots y_{k-n+1}^{0}\, y_{k-n+1}^{0})_b \\ y_k &:= y_{k-227} \text{ xor } \mathrm{shr}(h_k, 1) \text{ xor } ((h_k \bmod 2) \cdot \text{9908B0DF}) \end{aligned}$$

Um die Gleichverteilung sicher zu stellen modifiziert man die so generierte Zahlenfolge (y_k) ist aber noch - erst die Zahlenfolge (z_k) wird als Zufallszahlen verwendet. Für $k \geq 625$ berechnet man z_k durch:

$$\begin{aligned} x &:= y_k \text{ xor } \mathrm{shr}(y_k, 11) \\ y &:= x \text{ xor } (\mathrm{shl}(x, 7) \text{ and } \text{9D2C560}) \\ z &:= y \text{ xor } (\mathrm{shl}(y, 15) \text{ and } \text{EFC6000}) \\ z_k &:= z \text{ xor } \mathrm{shr}(z, 18) \end{aligned}$$

Kapitel 3

Kombinatorik

Die Kombinatorik ist das Grundgerüst der endlichen Wahrscheinlichkeitsrechnung. Hierbei stellt man sich immer die Frage wie viele Möglichkeiten es gibt eine bestimmte Situation zu realisieren. Denn je mehr Wege zu einem Ort führen, desto höher ist die Wahrscheinlichkeit diesen Ort zu finden. Die Frage, wie viele verschiedene Elementarereignisse es gibt, ist also immer eine kombinatorische Frage. Doch um die verschiedenen Möglichkeiten sinnvoll abzuzählen brauchen wir zunächst wieder ein bischen Notation:

Wir betrachten eine Menge mit $1 \leq n \in \mathbb{N}$ Elementen. Wenn wir eine lineare Ordnung auf dieser Menge einführen wollen, müssen wir ein kleinstes Element wählen. Im ersten Schritt gibt es dabei noch alle n Möglichkeiten. Nachdem das erste gewählt wurde, müssen wir das nächst größere Element wählen - wobei wir jetzt $n-1$ Möglichkeiten zur Auswahl haben. Fährt man immer so fort, so ergeben sich $n!$ Anordungen der n Elemente. Dabei ist die **Fakultät** $n!$ von n definiert, durch:

$$\begin{aligned} n! &:= \#\{\,\sigma : 1\ldots n \to 1\ldots n \mid \sigma \text{ bijektiv}\,\} \\ &= n\cdot(n-1)\cdot\ldots\cdot 2\cdot 1 \end{aligned}$$

Speziell für $n = 0$ setzen wir $0! := 1$. Als nächstes betrachten wir wieder unsere Menge $1\ldots n$ mit n Elementen. Wir wollen daraus eine Teilmenge I auswählen, die genau k Elemente enthält, wobei $k \in 0\ldots n$. Dazu müssen wir wieder mit einem ersten Element beginnen, wobei zunächst n Möglichkeiten bestehen. Im nächsten Schritt wählen wir wieder aus $n-1$ Elementen und so fort. Doch da die Reihenfolge der Elemente in I keine Rolle spielt, liefern je $k!$ verschiedene Ziehungen dieselbe Teilmenge I. Insgesamt finden wir die **Binomialkoeffizienten**

$$\begin{aligned} \binom{n}{k} &:= \#\{\,I \subseteq 1\ldots n \mid \#I = k\,\} \\ &= \frac{n}{k}\cdot\frac{n-1}{k-1}\cdots\frac{n-k+1}{1} \\ &= \frac{n!}{k!\,(n-k)!} \end{aligned}$$

Man beachte, dass diese Definition auch den Fall $k = 0$ einschließt, es ist $\binom{n}{0} = 1 = \binom{n}{n}$. Des weiteren erfüllen die Binomialkoeffizienten eine berühmte Rekursionsformel, die der Grund für den binomischen Lehrsatz (a und b können Elemente eines beliebigen Ringes sein, für die $ab = ba$ gilt) ist:

$$\begin{aligned} \binom{n+1}{k} &= \binom{n}{k} + \binom{n}{k-1} \\ (a+b)^n &= \sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k} \end{aligned}$$

Dies kann verallgemeinert werden: es seien a_1 bis $a_k \in R$ Elemente irgendeines kommutativen Ringes R und $a = (a_1, \ldots, a_k) \in R^k$. Weiterhin betrachten wir einen sogenannten *Multi-index* $\alpha = (\alpha_1, \ldots, \alpha_k) \in \mathbb{N}^k$. Dann führen wir die folgenden Notationen ein (wobei $n := |\alpha|$)

$$\begin{aligned} |\alpha| &:= \alpha_1 + \cdots + \alpha_k \\ \alpha! &:= (\alpha_1!) \cdot \ldots \cdot (\alpha_k!) \\ \binom{n}{\alpha} &:= \frac{n!}{\alpha!} \\ a^\alpha &:= (a_1)^{\alpha_1} \cdot \ldots \cdot (a_k)^{\alpha_k} \end{aligned}$$

Sind diese Bezeichnungen fixiert, dann gilt der polynomiale Lehrsatz (wobei die Summe über alle Multi-indices $\alpha \in \mathbb{N}^k$ mit $|\alpha| = \alpha_1 + \ldots + \alpha_k = n$ läuft)

$$(a_1 + \ldots + a_k)^n = \sum_{|\alpha|=n} \binom{n}{\alpha} a^\alpha$$

Eine Beweisskizze findet man bei [Scheja, Storch, Lehrbuch der Algebra §14] und einen Beweis bei [Zeidler, Abstract Algebra, Proposition 1.37]. Nachdem wir die Fakultät und die Binomialkoeffizienten eingeführt haben, betrachten wir der Reihe nach die verschiedenen kombinatorischen **Grundprobleme**:

1. **Schalterstellungen:** Wir haben n Schalter vorliegen. Jeder dieser Schalter besitzt k mögliche Stellungen. Wie viele Schalterstellungen sind dann möglich? Die Antwort liegt auf der Hand: n^k viele. Dieses Problem kann auch so formuliert werden: Wie viele Möglichkeiten gibt es Worte mit k Buchstaben Länge zu bilden, wenn das Alphabet n verschiedene Zeichen kennt? Wieder n^k. Noch eine Formulierung desselben Problems: Wir haben eine Urne mit n nummerierten Kugeln. Aus dieser ziehen wir k Mal, notieren die Nummer und legen die Kugel zurück. Wie viele verschiedene Ziehungen gibt es? Wieder n^k.

2. **Lottoziehung:** Wir haben n nummerierte Kugeln in einer Urne. Aus dieser ziehen wir k Mal und legen die Kugeln nicht wieder zurück (sondern in der Reihenfolge der Ziehung vor uns hin). Wie viele mögliche Ziehungen gibt es?

$$n(n-1)(n-2)\ldots(n-k+1) = \frac{n!}{(n-k)!} = k!\binom{n}{k}$$

Ein wichtiger Spezialfall ist der Folgende: Wollen wir die n Kugeln umordnen, so können wir dies tun, indem wir eine nach der anderen ziehen, bis alle n Stück gezogen wurden. D.h. es ist $k = n$ und damit gibt es $n!$ mögliche Anordnungen (**Permutationen**) der n Kugeln.

3. **Lottoauswertung:** Beim Lotto werden die Kugeln danach entsprechend der Reihenfolge ihrer Nummern sortiert, d.h. die Reihenfolge der Ziehung geht verloren. Je $k!$ verschiedene Lottoziehungen ergeben also ein und dasselbe Lottoergebnis. Entsprechend finden wir die folgende Zahl möglicher Lottoergebnisse für k aus n Kugeln:

$$\binom{n}{k} = \frac{n}{1} \cdot \frac{n-1}{2} \cdot \ldots \cdot \frac{n-k+1}{k}$$

Dieses Problem lässt sich auch so formulieren: Wir haben n nummerierte Kugel in Urne 1. Aus dieser Urne 1 ziehen wir k Mal (ohne Zurücklegen) und legen die gezogene Kugel in eine zweite Urne 2. Dann gibt es wieder $\binom{n}{k}$ Möglichkeiten Urne 2 zu füllen.

4. **Zweikugelordnung:** Wir haben N Kugeln gegeben, n weiße und $m = N - n$ schwarze. Wie viele Arten gibt es diese Kugeln anzuordnen (d.h. in eine Reihe zu legen)? Die Antwort lautet:

$$\binom{N}{n} = \binom{n+m}{n}$$

Begründung: wir schreiben die Nummern 1 bis N auf die Kugeln. Es gibt $N!$ Möglichkeiten die nummerierten Kugeln anzuordnen. Wischen wir die Nummern jetzt wieder weg, so spielt es aber keine Rolle mehr, in welcher Reihenfolge die weißen (bzw. die schwarzen) Kugeln untereinander liegen. Je $n!$ für die weißen und $(N-n)!$ für die schwarzen Anordnungen sind also identisch. D.h. die $N!$ Anordnungen der nummerierten Kugeln reduzieren sich auf $N!/(n!(N-n)!) = \binom{N}{n}$ Anordnungen für die unnummerierten Kugeln.

5. **Kugelverteilung I:** Wir haben n nummerierte Kugeln und k Urnen vorliegen. Wir wollen nun die Kugeln so auf die Urnen verteilen, dass in die erste Urne α_1 Kugeln kommen, und ... und in die k-te Urne α_k Kugeln kommen ($n = |\alpha| = \alpha_1 + \alpha_2 + \cdots + \alpha_k$). Wie viele Möglichkeiten gibt es das zu tun?

$$\binom{n}{\alpha} = \frac{n!}{\alpha_1!\,\alpha_2! \cdot \ldots \cdot \alpha_k!}$$

Begründung: Es gibt $n!$ Möglichkeiten die n Kugeln anzuordnen. Die ersten α_1 kommen in Urne $1, \ldots,$ die letzten α_k kommen in Urne k. Die Anordnung innerhalb der Urnen ist aber egal. D.h. wir müssen durch $\alpha_1!$ (für Urne 1), und ... und durch $\alpha_k!$ (für Urne k) dividieren.

6. **Kugelverteilung II:** Wir haben n identische Kugeln und k Urnen gegeben. Wir wollen diese Kugeln irgendwie auf die Urnen verteilen. Wie viele verschiedene Verteilungen gibt es?

$$\binom{n+k-1}{n}$$

Begründung: Wir führen eine günstige Art ein, die Verteilungen zu codieren: eine weiße Kugel $\circ$ codiert eine der n Kugeln, und eine schwarze Kugel $\bullet$ codiert eine Trennwand. Zum Beispiel kodiert $\circ\ \circ\ \bullet\ \bullet\ \circ$ folgende Situation: es gibt 3 Kugeln und 3 Urnen. Die erste Urne enthält 2 Kugeln, die zweite Urne ist leer und die dritte Urne enthält eine Kugel. Wir haben also n weiße Kugeln und (bei k Urnen) $m = k-1$ schwarze Kugeln zur Codierung. Nach (4) gibt es genau $\binom{n+m}{n}$ solche Codes, also $\binom{n+k-1}{n}$ Verteilungen.

7. **Häufigkeitsverteilungen:** Wir haben eine Urne mit n nummerierten Kugeln. Aus dieser ziehen wir k Mal (mit Zurücklegen) und notieren die Häufigkeiten, wie oft welche Nummer gezogen wurde. Wie viele Häufigkeitsverteilungen gibt es?

$$\binom{n+k-1}{k}$$

Begründung: Zu jeder Kugelnummer nehmen wir eine Urne. Und wir stellen einen Sack mit k gleichartigen Bällen bereit. Wir ziehen nun eine Kugel aus der Urne, legen sie wieder zurück und tun dafür einen Ball aus dem Sack in die Urne der gezogenen Kugel. Das Ganze machen wir k Mal. Die Zahl der Bälle in den Urnen gibt also die Häufigkeitsverteilung an. Dazu haben wir aber k identische Bälle auf n Urnen verteilt. Und nach (6) gibt es dabei $\binom{k+n-1}{k}$ Möglichkeiten.

Bemerkung 3.1: Stirling'sche Formeln:
Für große Werte von n lässt sich die Fakultät $n!$ recht gut annähern, durch:

$$n! \;=\; (1+\varepsilon_n)\left(\frac{n}{e}\right)^n \sqrt{2\pi n} \quad \text{mit} \quad 0 < \varepsilon_n < \frac{1}{11n}$$

Wendet man diese Formel auf den Binomialkoeffizienten an, so erhält man für große Werte von n und kleine Werte von k die folgende Formel:

$$\binom{n}{k} \;=\; (1+\delta_{n-k})\frac{1}{e^k \cdot k!} \cdot \frac{n^n}{(n-k)^{n-k}} \cdot \sqrt{\frac{n}{n-k}} \quad \text{mit} \quad \frac{-1}{11x} < \delta_x < \frac{1}{11x}$$

Sind hingegen sowohl n, als auch k groß, so kann man auch $k!$ annähern zu:

$$\binom{n}{k} \;=\; (1+\Delta_{n,k})\frac{n^n}{k^k \cdot (n-k)^{n-k}} \cdot \sqrt{\frac{n}{2\pi k(n-k)}}$$

$$\text{mit} \quad -\frac{n+0.1}{11k(n-k)+n} \;<\; \Delta_{k,n} \;<\; \frac{1}{11n}$$

Beispiel 3.2: Geburtstagsproblem:
Als Beispiel eines kombinatorischen Problems stellen wir uns folgende Frage: Wie hoch ist die Wahrscheinlichkeit, dass unter einer Gruppe von n Personen mindestens 2 am selben Tag (Datum ohne Jahr) Geburtstag haben? Sei p die gesuchte Wahrscheinlichkeit, dann ist $q = 1 - p$ also die Wahrscheinlichkeit des Ereignisses: alle n Personen haben an verschiedenen Tagen Geburtstag.

Der Einfachheit halber ignorieren wir Schaltjahre, und nehmen an jeder Tag komme gleich häufig als Geburtstag vor. Dann ist dies ist eine äquivalente Formulierung der Lottoziehung (2) oben: Sei $T = 365$ die Anzahl der Tage eines Jahres. Für die erste Person gibt es T Möglichkeiten, wann sie Geburtstag hat, für die zweite Person bleiben noch $T - 1$ Möglichkeiten, da eine ja schon vergeben ist, und so weiter. Da jeder Tag mit Wahrscheinlichkeit $1/T$ als Geburtstag vorkommt, haben wir also die folgende Wahrscheinlichkeit q

$$q \;=\; \frac{T}{T} \cdots \frac{T-1}{T} \cdot \cdots \cdot \frac{T-n+1}{T} \;=\; \frac{T!}{T^n (T-n)!} \;=\; \frac{n!}{T^n} \binom{T}{n}$$

Und wendet man die Stirling'sche Formel auf $T!$ und $(T-n)!$ an, dann lässt sich die gesuchte Wahrscheinlichkeit p recht genau annähern, durch:

$$p \;\approx\; 1 - \left(\frac{T}{T-n} \right)^{T+0.5-n}$$

Kapitel 4

Eindimensionale Statistik

Beispiel 4.1:
Ein (sechsseitiger) Würfel wird 50 mal geworfen und die Ergebnisse - welche Zahl wie oft gefallen ist - notiert. Es ergibt sich folgende Tabelle:

Augenzahl	Anzahl Würfe	relative Häufigkeit
1	9	18 %
2	10	20 %
3	8	16 %
4	7	14 %
5	7	14 %
6	9	18 %

Die geläufigste Kennzahl, die angibt, in welchem Bereich diese Werte liegen, ist der Durchschnitt - das arithmetische Mittel. Dieses berechnet sich zu

$$\frac{1}{50}\left(9 \cdot 1 + 10 \cdot 2 + 8 \cdot 3 + 7 \cdot 4 + 7 \cdot 5 + 9 \cdot 6\right) \quad = \quad 3.4$$

Das arithmetische Mittel sagt zwar um welche Zahl sich die Werte gruppieren, aber noch nichts über die Streuweite der Werte. Dies leistet eine andere Kennzahl: die Varianz. Für diese berechnet man den Durchschnitt der quadratischen Abweichungen der Werte von dem Mittelwert:

$$\frac{1}{50}\left(9 \cdot 2.4^2 + 10 \cdot 1.4^2 + 8 \cdot 0.4^2 + 7 \cdot 0.6^2 + 7 \cdot 1.6^2 + 9 \cdot 2.6^2\right) \quad = \quad 3.08$$

Durchschnitt und Varianz repräsentieren also die Position und die Breite dieser 50 Würfelergebnisse in kompakter Form. Bevor wir auf die geometrische Interpretation dieser Werte eingehen, sollten wir zunächst allgemein betrachten, was wir gerade getan haben:

Definition 4.2:
Sei S eine beliebige Menge, genannt der *Merkmalsraum*, und sei $1 \leq n \in \mathbb{N}$ eine natürliche Zahl. Dann nennen wir ein n-Tupel $x = (x_1, \ldots, x_n) \in S^n$ eine *Stichprobe* vom *Umfang* n. Und für ein $s \in S$ bezeichnen wir die *absoluten Häufigkeiten* $n(s)$ und *relativen Häufigkeiten* $h(s)$ von s in x:

$$\begin{aligned} n(s) &:= \#\{\, i \in 1 \ldots n \mid x_i = s \,\} \\ h(s) &:= \frac{n(s)}{n} \end{aligned}$$

Ist $S \subseteq \mathbb{R}$ eine Teilmenge der reellen Zahlen, dann definieren wir weiterhin das **arithmetische Mittel** $A(x)$, die **Varianz** $V(x)$ und die **Standardabweichung** $\sigma(x)$ der Stichprobe x durch:

$$\begin{aligned} A(x) &:= \frac{1}{n} \sum_{i=1}^{n} x_i \\ V(x) &:= \frac{1}{n} \sum_{i=1}^{n} (x_i - A(x))^2 \\ \sigma(x) &:= \sqrt{V(x)} \end{aligned}$$

Bemerkung: In dem obigen Beispiel ist also $S = 1 \ldots 6$ die Menge der möglichen Würfelergebnisse, x sind die Würfelergebnisse (in Reihenfolge) und die Tabelle gibt die absoluten und relativen Häufigkeiten in x an. Offenbar genügen diese für die Berechnung des arithmetischen Mittels und der Varianz.

Proposition 4.3:
Sei $1 \leq n \in \mathbb{N}$ eine natürliche Zahl und $x \in \mathbb{R}^n$ eine Stichprobe vom Umfang n über $\mathbb{R}$. Wir verwenden die Bezeichnungen aus (4.2) und Konventionen, wie sie in (2.2) gegeben wurden. Dann gilt:

(i) Das arithmetische Mittel $A : \mathbb{R}^n \to \mathbb{R} : x \mapsto A(x)$ ist eine lineare Abbildung und ist $a \in \mathbb{R}$ eine reelle Zahl, so gilt weiterhin

$$\begin{aligned} A(a) &= a \\ V(a) &= 0 \\ V(x+a) &= V(x) \end{aligned}$$

(ii) Bezeichnet $S := \{\, x_i \mid i \in 1 \ldots n \,\} \subseteq \mathbb{R}$ die (endliche) Menge der *verschiedenen* Werte, die in der Stichprobe x vorkommen, dann können das arithmetische Mittel und die Varianz über die relativen Häufigkeiten berechnet werden, vermöge:

$$\begin{aligned} A(x) &= \sum_{s \in S} h(s) s \\ V(x) &= \sum_{s \in S} h(s) s^2 - A(x)^2 \;=\; A(x^2) - A(x)^2 \end{aligned}$$

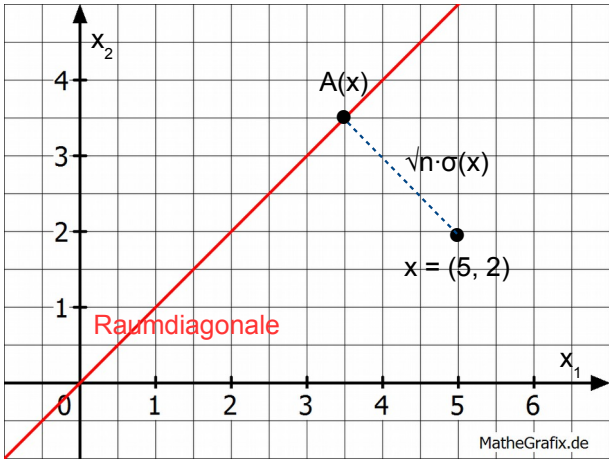
(iii) Das arithmetische Mittel ist die beste Approximation von $x \in \mathbb{R}^n$ durch eine einzige Zahl $\overline{x} \in \mathbb{R}$ gemessen in der natürlichen Norm:

$$\|x - A(x)\| \;=\; \min\{\, \|x - \overline{x}\| \mid \overline{x} \in \mathbb{R} \,\}$$

Die Standardabweichung misst dabei den Abstand von $x \in \mathbb{R}^n$ zu dem arithmetischen Mittel $A(x) \in \mathbb{R}^n$ auf der Raumdiagonalen:

$$\sigma(x) \;=\; \frac{1}{\sqrt{n}}\, \|x - A(x)\|$$

Bemerkung 4.4:
Die Eigenschaft (4.3.iii) des arithmetischen Mittels besagt also, dass $A(x)$ gerade die orthogonale Projektion des Punktes $x \in \mathbb{R}^n$ auf die Raumdiagonale $\Delta(t) := t\mathbf{1}$ (mit $\mathbf{1} := (1, 1, \ldots, 1) \in \mathbb{R}^n$) ist. Die folgende Abbildung illustriert dies im Fall von $n = 2$ Dimensionen:



Dies liegt einfach daran, dass das arithmetische Mittel definiert wurde, als $A(x) = (x \bullet \mathbf{1})/(\mathbf{1} \bullet \mathbf{1})$. Und nach unserer Konvention wird die *Zahl* $A(x)$ ja als Punkt $A(x)\mathbf{1}$ auf der Raumdiagonalen interpretiert. Hier sei nur erwähnt, dass der Median $Z(x)$ ebenfalls eine orthogonale Projektion ist: Dabei wird der Abstand aber nicht in der 2- sondern in der 1-Norm gemessen. Wollte man die absolute Abweichung minimieren, wäre das zugehörige Lagemaß nicht der Durchschnitt, sondern der Median. Näheres dazu findet sich in [Zeidler, Algebraic Statistics, auf www.wuala.com/zeidlerweb].

Definition 4.5:
Sei wieder $S = \{ s_1, s_2, \ldots, s_m \}$ eine endliche Menge und $\leq$ eine totale Ordnung auf S. Die Nummerierung der s_p sei so gewählt, dass gilt

$$s_1 \ < \ s_2 \ < \ \ldots \ < \ s_m$$

Sei weiterhin $x \in S^n$ für einen Stichprobenumfang $1 \leq n \in \mathbb{N}$. Der Einfachheit halber schreiben wir $n_p := n(s_p)$ bzw. $h_p := h(s_p)$ für die absoluten bzw. relativen Häufigkeiten. Dann definieren wir weiterhin die **kumulierten absoluten** N_q bzw. **relativen Häufigkeiten** H_q durch

$$\begin{aligned} N_q &:= \sum_{p=1}^{q} n_q \\ H_q &:= \sum_{p=1}^{q} h_q \ = \ \frac{1}{n} N_q \end{aligned}$$

Der Vollständigkeit halber setzen wir auch $N_0 := 0$ und $H_0 := 0$. Sei nun $j \in 1 \ldots 3$, dann definieren wir das j-te **Quartil** von x als denjenigen Wert s_p für den die kumulierte, relative Häufigkeit erstmals über $j/4$ steigt:

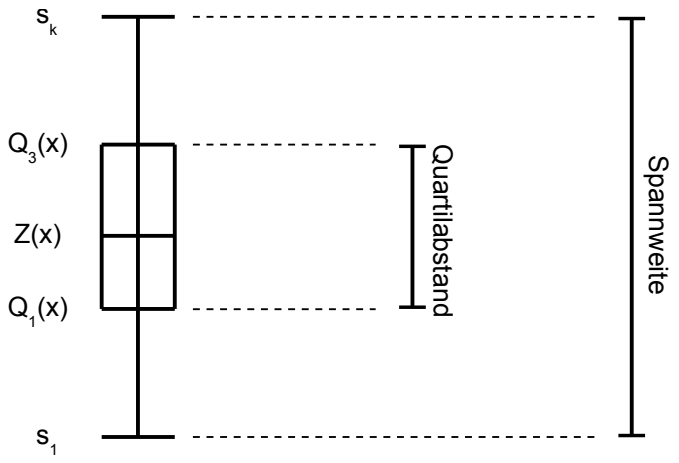
$$Q_j(x) \ := \ s_p \quad \text{wobei} \quad p \ = \ \min\left\{ q \in 1 \ldots m \mid \frac{j}{4} \leq H_q \right\}$$

Das zweite Quartil wird auch Zentralwert oder **Median** genannt, wir schreiben dafür auch $Z(x) := Q_2(x)$. Würde man die x_k der Reihe nach (aufsteigend) sortieren, so wäre $Z(x)$ der Wert in der Mitte (links).

Beispiel 4.6:
Wir betrachten wieder unsere Würfelwürfe zu Beginn des Kapitels. Hierbei ist die p-te Augenzahl $s_p = p$ gerade die Zahl p selbst. Wir ergänzen die Tabelle also um die kumulierten Häufigkeiten:

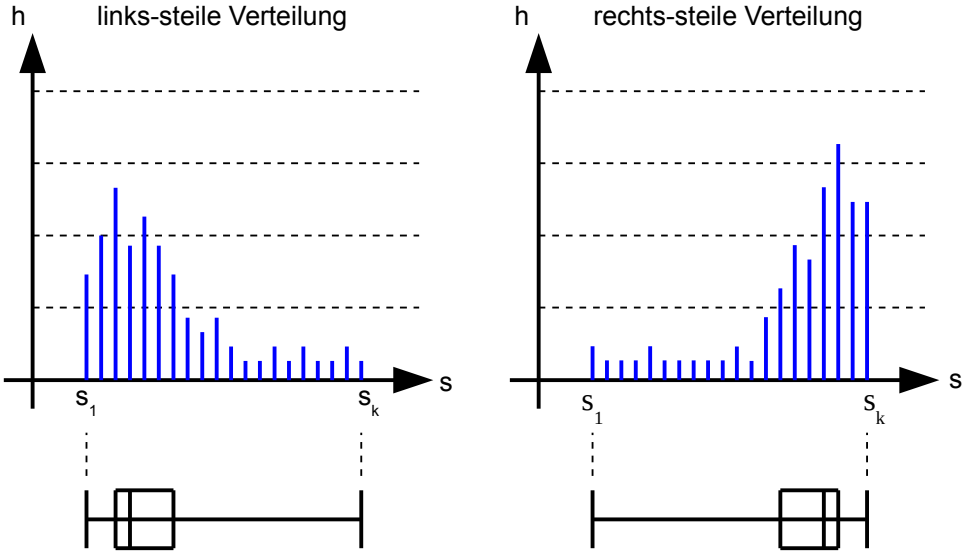
s_p	n_p	N_p	h_p	H_p
1	9	9	0.18	0.18
2	10	19	0.20	0.38
3	8	27	0.16	0.54
4	7	34	0.14	0.68
5	7	41	0.14	0.82
6	9	50	0.18	1.00

Wir sehen, dass 0.25 erstmals für $s_p = 2$ erreicht wird, es ist also $Q_1(x) = 2$. Genauso wird 0.50 erstmals für $s_p = 3$ und 0.75 erstmals für $p = 5$ überschritten. Es sind also $Z(x) = 3$ und $Q_3(x) = 5$. Die Quartile lassen sich recht übersichtlich in einem **Box-Diagramm** darstellen:



Bemerkung 4.7:
Sind die niedrigen Werte häufiger, als große Werte, spricht man auch von einer *links-steilen* Verteilung. Ein Beispiel dafür ist etwa die Vermögensverteilung in Deutschland, oder - noch extremer - auf der Welt. Überwiegen hingegen die hohen Werte, spricht man von einer *rechts-steilen Verteilung*. Welche Art von Verteilung vorliegt, erkennt man anhand der Lage der Quartile zueinander und zum arithmetischen Mittel:

links-steile Verteilung	rechts-steile Verteilung
$Z(x) < A(x)$	$A(x) < Z(x)$
$Z(x) < \frac{1}{2}(Q_1(x) + Q_3(x))$	$\frac{1}{2}(Q_1(x) + Q_3(x)) < Z(x)$



Wir wollen nun ein Maß für die Ungleichheit einer Verteilung einführen. Das bekannteste Maß dieser Eigenschaft ist der *Gini-Koeffizient*. Dieser wird zwar nicht an der Schule unterrichtet, ist aber insbesondere bei Wirtschaftswissenschaftlern sehr beliebt. Wer Statistik betreiben möchte, kommt also an diesem Begriff nicht vorbei, so dass wir hier eine Konstruktion dieser Kennzahl geben möchten:

Definition 4.8: Konzentrationsanalyse:
Sei $x \in \mathbb{R}^n$ ein Datensatz bestehend aus positiven Zahlen $x_k \geq 0$. Wir bezeichnen die *verschiedenen* Werte unter den x_k mit s_i, d.h.

$$S = \{ s_1, s_2, \ldots, s_m \} := \{ x_1, x_2, \ldots, x_n \} \subseteq \mathbb{R}^+$$

Wir nehmen an die x_k und die s_i seien aufsteigend sortiert, d.h. es gelte $x_1 \leq x_2 \leq \cdots \leq x_n$ und $s_1 < s_2 < \cdots < s_m$. Dann definieren wir die folgenden Größen (für $j \in 1 \ldots m$):

$$\begin{aligned}
X &:= nA(x) = \sum_{k=1}^{n} x_k = \sum_{i=1}^{m} n(s_i)s_i \\
\ell(s_i) &:= \frac{n(s_i)s_i}{X} = \frac{h(s_i)s_i}{A(x)} \\
L_j &:= \sum_{i=1}^{j} \ell(s_i) = \frac{1}{X}\sum_{i=1}^{j} n(s_i)s_i \\
H_j &:= \sum_{i=1}^{j} h(s_i) = \frac{1}{n}\sum_{i=1}^{j} n(s_i)
\end{aligned}$$

Speziell für $j = 0$ bezeichnen wir noch $L_0 := 0$ und $H_0 := 0$. Dabei nennen wir die H_j die *kumulierten relativen Häufigkeiten* der s_j. Dabei gilt stets $0 \leq L_j \leq H_j \leq 1$. Verbinden wir die Punkte $(H_j \mid L_j)$ nun im Koordinatensystem so entsteht die **Lorenzkurve** L von x. Diese ist also eine Abbildung der Form $L : [0,1] \to [0,1]$, gegeben durch

$$L(H) = \frac{\ell(s_i)}{h(s_i)}(H - H_{i-1}) + L_{i-1} \quad \text{wobei } H_{i-1} \leq H \leq H_i$$

Damit ist $L(H)$ eine Kurve, die unter der Diagonalen $H \mapsto H$ durchhängt. Die Größe des nach unten hängenden Bauchs der Kurve nennen wir den **Gini-Koeffizienten** $G(x)$ von x:

$$\begin{aligned}
G(x) &:= 2\int_0^1 H - L(H)dH \\
&= 1 - \sum_{i=1}^{m} h_i(L_i + L_{i-1}) \\
&= \frac{2}{nX}\sum_{k=1}^{n} kx_k - \frac{n+1}{n}
\end{aligned}$$

Da der Bauch aber maximal die Fläche 1/2 annehmen kann, haben wir die Fläche noch mit 2 multipliziert, um eine Kennzahl $G(x) \in [0,1]$ zu erhalten. Dabei interpretieren wir Werte von $G(x) = 0$ bis 0.25 als gute Gleichverteilung und Werte von $G(x) = 0.4$ bis 1 als starke Ungleichverteilung.

Beispiel 4.9:

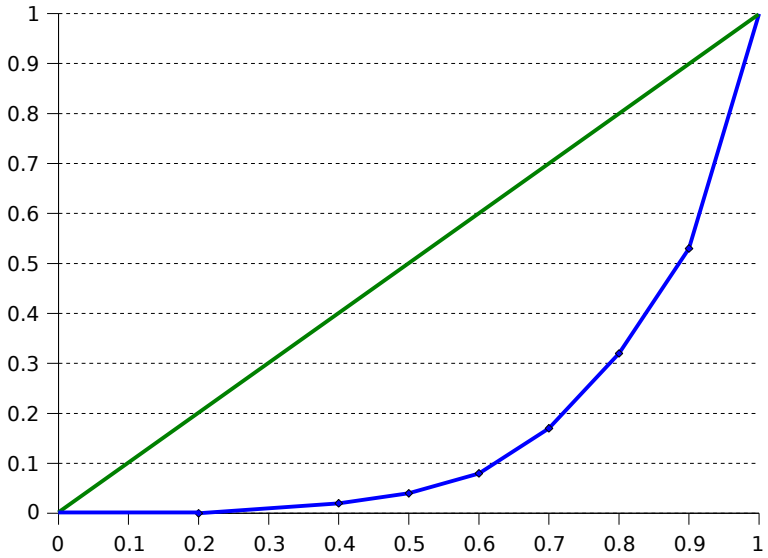
(i) Haben alle gleich viel - also $e(n) := (a, a, \ldots, a) \in \mathbb{R}^n$ für irgendeine positive Zahl $a > 0$, dann beträgt der Gini-Koeffizient $G(e(n)) = 0$.

(ii) Hat einer alles, also $s(n) := (0, 0, \ldots, 0, a) \in \mathbb{R}^n$ wobei wieder $a > 0$ beliebig ist, dann nimmt der Gini-Koeffizient den Wert $G(s(n)) = 1 - 1/n$ an. Insbesondere gilt also $G(s(n)) \to 1$ für $n \to \infty$.

(iii) Angenommen in Deutschland gäbe es 10 Haushalte und das Gesamtvermögen der Haushalte sei auf 100 Einheiten normiert. Dann sähe die Vermögensverteilung in Deutschland etwa folgendermaßen aus:

$$0 \leq 0 \leq 1 \leq 1 \leq 2 \leq 4 \leq 9 \leq 15 \leq 21 \leq 47$$

Wir haben also $n = 10$, $m = 8$ und die Zahlenpaare $(H_i \mid L_i)$ betragen hintereinander weg: $(0.2 \mid 0)$, $(0.4 \mid 0.02)$, $(0.5 \mid 0.04)$, $(0.6 \mid 0.08)$, $(0.7 \mid 0.17)$, $(0.8 \mid 0.32)$, $(0.9 \mid 0.53)$ und $(1, 1)$. Die Lorenzkurve demonstriert die Ungleichverteilung eindrucksvoll:



Lorenz-Kurve (unten) der Vermögensverteilung in Deutschland, die Fläche zwischen den Kurven ist der halbe Gini-Koeffizient G

Der Gini-Koeffizient dieser Verteilung berechnet sich damit zu $G = 0.68$ und obwohl dieser Wert damit eine sehr starke Ungleichverteilung anzeigt, gehört Deutschland damit noch zu den Nationen mit eher gemäßigter Ungleichverteilung.

(iv) Wir wählen zwei Zahlen $a \leq b \in \mathbb{R}$ und betrachten die Binomialverteilung mit $1 \leq m \in \mathbb{N}$ Stufen. D.h. wir betrachten die Daten $x \in \mathbb{R}^n$ (mit $n = 2^m$) bestehend aus den verschiedenen Werten $S = \{ s_i \mid i \in 0 \ldots m \}$ wobei $s_i := a^{m-i}b^i$ genau n_i mal in x vorkommt, mit $n_i = \binom{m}{i}$.

$$\begin{aligned}
X &= \sum_{k=1}^{n} x_k = \sum_{i=0}^{m} n_i s_i = (a+b)^m \\
h_i &= \frac{n_i}{n} = \frac{\binom{m}{i}}{2^m} \\
H_j &= \sum_{i=0}^{j} h_i = \frac{1}{2^m} \sum_{i=0}^{j} \binom{m}{i} \\
\ell_i &= \frac{n_i s_i}{X} = \frac{\binom{m}{i} a^{m-i} b^i}{(a+b)^m} \\
L_j &= \sum_{i=0}^{j} \ell_i = \frac{1}{(a+b)^m} \sum_{i=0}^{j} \binom{m}{i} a^{m-i} b^i
\end{aligned}$$

Man beachte, dass im Gegensatz zur Definition oben die Zählung bei 0 beginnt, d.h. der erste Knick der Lorenzkurve liegt bei $(L_0 \mid H_0)$. Eine kurze Rechnung zeigt dann, dass sich der Gini-Koeffizient dieser Verteilung berechnen lässt, durch:

$$G(x) = \frac{1}{2^m} \sum_{j=0}^{m} \left(\sum_{0 \leq i < j} d_i + \frac{d_j}{2} \right)$$

$$d_i := h_i - \ell_i = \frac{\binom{m}{i}}{(2a+2b)^m} \left((a+b)^m - 2^m a^{m-i} b^i \right)$$

Satz 4.10:
Sei $x \in \mathbb{R}^n$ eine sortierte Liste von positiven Zahlen $0 \leq x_1 \leq \cdots \leq x_n$ und $a > 0$ eine echt positive Zahl. Dann ist der Gini-Koeffizient von x invariant unter der Streckung mit a und wird bei der Verschiebung um a verringert:

$$\begin{aligned}
G(ax) &= G(x) \\
G(x+a) &= G(x) - \frac{a}{X(X+na)} \sum_{k=1}^{n} (n+1-k) x_k
\end{aligned}$$

Kapitel 5

Zweidimensionale Statistik

Bisher haben wir nur eindimensionale Daten betrachtet - wo sie liegen und wie weit sie streuen und wie Ungleich sie verteilt sind. Oft sind aber gerade die Zusammenhänge zwischen Daten interessant. In diesem Kapitel wollen wir also Methoden vorstellen, wie man den Zusammenhang zwischen zwei verschiedenen Merkmalen herausarbeiten kann. Als Beispiel betrachten wir den Zusammenhang zwischen beruflicher Ausbildung und Beschäftigungsverhältnis. Eine Umfrage in Deutschland im Jahre 2013 unter 10000 repräsentativ ausgewählten Bürgern im Alter von 15 oder mehr Jahren hätte in etwa folgendes Ergebnis gezeigt:

Ausbildung	arbeitslos	atypisch	angestellt	selbstständig	Summe
ohne	312	977	1216	178	2683
Lehre	330	610	4303	656	5899
Studium	46	447	725	200	1418
Summe	688	2034	6244	1034	10000

Ein *atypisches* Beschäftigungsverhältnis ist befristet, Leiharbeit, in Teilzeit oder geringfügig. Im Gegensatz dazu bezeichnet *angestellt* ein festes Anstellungsverhältnis. Die absoluten Zahlen sind immer etwas schwer zu fassen, deswegen ist es leichter die relativen Häufigkeiten zu analysieren:

Ausbildung	arbeitslos	atypisch	angestellt	selbstständig	Summe
ohne	3.1 %	9.8 %	12.2 %	1.8 %	26.8 %
Lehre	3.3 %	6.1 %	43.0 %	6.6 %	59.0 %
Studium	0.5 %	4.5 %	7.3 %	2.0 %	14.2 %
Summe	6.9 %	20.3 %	62.5 %	10.3 %	100 %

Hier sieht man sofort die Arbeitslosenquote, die 2013 bei nur 6.9% lag (weil diese Zahl die vielen atypischen Beschäftigungsverhältnisse versteckt). Interessant ist es damit beispielsweise die Arbeitslosenquote der Akademiker zu vergleichen. Diese lag bei nur

$$h(\text{arbeitslos} \mid \text{Studium}) \;=\; \frac{46}{1418} \;\approx\; 3.2\%$$

Ein weiteres Beispiel: 26.8% der Befragten hatten keine abgeschlossene Berufsausbildung (dies würde ähnlich auch für ein Alter ab 25 gelten)! Unter den atypisch Beschäftigten ist die Quote derer ohne Berufsausbildung aber

noch deutlich höher:

$$h(\text{ohne Berufsausbildung} \mid \text{atypisch}) \;=\; \frac{977}{2034} \;\approx\; 48.0\%$$

Definition 5.1:
Es seien $S = \{ s_1, \ldots, s_u \}$ und $T = \{ t_1, \ldots, t_v \}$ zwei beliebige endliche Mengen. Sei ferner $x \in (S \times T)^n$ eine Stichprobe vom Umfang $1 \leq n \in \mathbb{N}$ aus dem Merkmalsraum $S \times T$, dann bezeichnen wir die *absolute* und die *relative Häufigkeit* $n(s,t)$ bzw. $h(s,t)$ des Paares $(s,t) \in S \times T$ durch

$$\begin{aligned} n(s,t) &:= \#\{ i \in 1 \ldots n \mid x_i = (s,t) \} \\ h(s,t) &:= \frac{n(s,t)}{n} \end{aligned}$$

Für $s \in S$ und $t \in T$ bezeichnen wir zudem die *absoluten* und *relativen Randhäufigkeiten* der Stichprobe x, mit

$$\begin{aligned} n(s,+) &:= \sum_{t \in T} n(s,t) \\ n(+,t) &:= \sum_{s \in S} n(s,t) \\ h(s,+) &:= \sum_{t \in T} h(s,t) \;=\; \frac{n(s,+)}{n} \\ h(+,t) &:= \sum_{s \in S} h(s,t) \;=\; \frac{n(+,t)}{n} \end{aligned}$$

Unter der *bedingten Häufigkeit* von $s \in S$ unter der Bedingung $t \in T$ (und umgekehrt) verstehen wir die folgenden Quotienten (sofern möglich):

$$\begin{aligned} h(s \mid t) &:= \frac{n(s,t)}{n(+,t)} \;=\; \frac{h(s,t)}{h(+,t)} \\ h(t \mid s) &:= \frac{n(s,t)}{n(s,+)} \;=\; \frac{h(s,t)}{h(s,+)} \end{aligned}$$

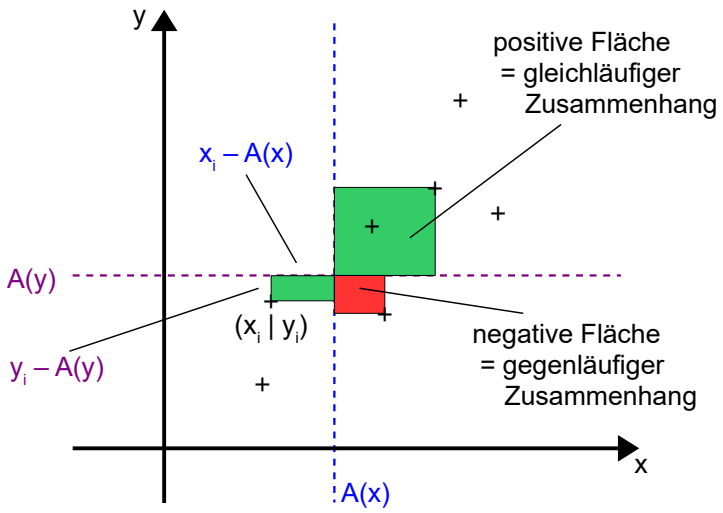
Als Maß für die Unabhängigkeit der Werte definiert man die **mittlere quadratische Kontingenz** $C(x)$ und die **korrigierte Kontingenz** $C^*(x)$ der Stichprobe x, durch (wobei $\mu := \min\{ u, v \}$)

$$\begin{aligned} C(x) &:= \sum_{s \in S} \sum_{t \in T} \frac{(h(s,t) - h(s,+)h(+,t))^2}{h(s,+)h(+,t)} \\ C^*(x) &:= \sqrt{\frac{\mu}{\mu - 1}} \cdot \sqrt{\frac{C(x)}{C(x)+1}} \end{aligned}$$

Auf den Begriff der Unabhängigkeit zweier Größen werden wir später noch genauer eingehen. Da wir aber zumeist mit numerischen Größen zu tun haben, müssen wir uns nicht auf das Auswerten von Häufigkeiten beschränken, sondern können versuchen einen funktionalen Zusammenhang zwischen x und $y \in \mathbb{R}^n$ zu finden. Dazu dienen die folgenden Begriffe:

Definition 5.2:
Sind $x \in \mathbb{R}^n$ und $y \in \mathbb{R}^n$ zwei n-Tupel reeller Zahlen, dann definieren wir deren **Kovarianz** $\sigma(x \mid y)$ und **Korrelation** $r(x, y)$ durch

$$\begin{aligned} \sigma(x \mid y) &:= \frac{1}{n}\sum_{i=1}^{n}(x_i - A(x)) \cdot (y_i - A(y)) \\ r(x \mid y) &:= \frac{1}{n}\sum_{i=1}^{n}\frac{x_i - A(x)}{\sigma(x)} \cdot \frac{y_i - A(y)}{\sigma(y)} = \frac{\sigma(x \mid y)}{\sigma(x)\sigma(y)} \end{aligned}$$



Proposition 5.3:
Wir verwenden wieder die in (2.2) eingeführten Notationen. Seien x und $y \in \mathbb{R}^n$ zwei n-Tupel, wobei $1 \leq n \in \mathbb{N}$ und $a \in \mathbb{R}$ eine Zahl. Dann ist die Kovarianz $\sigma : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ eine symmetrische Bilinearform, für die gilt:

$$\begin{aligned} \sigma(a \mid y) &= 0 \\ \sigma(x \mid x) &= \sigma(x)^2 \\ \sigma(x \mid y) &= A(xy) - A(x)A(y) \end{aligned}$$

Wir nehmen nun weiterhin an, x und y seien *zentriert*, d.h. es gelte $A(x) = 0 = A(y)$. Ist nun eine $(n \times n)$-Matrix $T \in \mathrm{mat}_n(\mathbb{R})$ gegeben, dann gilt

$$\sigma(Tx \mid y) = \sigma(x \mid T^* y)$$

Insbesondere ist die Varianz zentrierter Variablen invariant unter unitären Transformationen. D.h. ist T unitär (das bedeutet $T^{-1} = T^*$), dann gilt:

$$V(Tx) = V(x)$$

Parameter-Fitting:
Eine der Hauptaufgaben der Statistik ist es die beste Abschätzung für Parameter in mathematischen Modellen realer Vorgänge zu liefern. Wenn man beispielsweise den Verlauf des Luftdrucks p in Abhängigkeit der Höhe h über dem Meeresspiegel beschreiben möchte, dann führt eine klassische Betrachtung zu einer Abhängigkeit der Form $p(h) = p_0 e^{-\alpha h}$ für irgend welche Parameter p_0 und α. Um diese Werte zu bestimmen, müsste man Messungen des Luftdrucks in verschiedenen Höhen durchführen. Wir haben also Meßwerte (h_i, p_i) zur Verfügung und sollen daraus p_0 und α so ermitteln, dass der unter diesen Werten berechneten theoretische Luftdruck $p_0 e^{-\alpha h_i}$ möglichst genau mit dem gemessenen Luftdruck p_i überein stimmt. Wir wollen uns nun also dieser Aufgabe stellen:

Proposition 5.4: Methode der kleinsten Quadrate
Es sei S eine beliebige Menge, $1 \leq m, n \in \mathbb{N}$ und wir betrachten weiterhin die Messwerte $x = (x_1, \ldots, x_n) \in S^n$ und $y = (y_1, \ldots, y_n) \in \mathbb{R}^n$. Sei schließlich noch zu jedem $i \in 1 \ldots m$ eine Funktion $f_i : S \to \mathbb{R}$ gegeben. Dann definieren wir die folgende $(n \times m)$-Matrix

$$F \;:=\; \begin{pmatrix} f_1(x_1) & \cdots & f_m(x_1) \\ \vdots & & \vdots \\ f_1(x_n) & \cdots & f_m(x_n) \end{pmatrix}$$

Die transponierte Matrix von F nennen wir F^*. Dann ist F^*F eine symmetrische $(m \times m)$-Matrix, die explizit geschrieben werden kann als

$$F^*F \;=\; \left(\sum_{k=1}^{n} f_i(x_k) f_j(x_k) \right)_{i,j \in 1 \ldots m}$$

Schließlich sei a ein m-Tupel $a = (a_1, a_2, \ldots, a_m) \in \mathbb{R}^m$, mit dem wir die *Näherungsfunktion* f_a und die *Abstandsfunktion* d definieren, durch

$$f_a \;:\; S \to \mathbb{R} \;:\; x \mapsto \sum_{i=1}^{m} a_i f_i(x)$$

$$d \;:\; \mathbb{R}^m \to \mathbb{R} \;:\; a \mapsto \sqrt{\sum_{k=1}^{n} (f_a(x_k) - y_k)^2}$$

Man beachte, dass dabei stets $f_a(x_k) = [Fa]_k$ ist, mit anderen Worten gilt $Fa = f_a(x) := (f_a(x_1), \ldots, f_a(x_n)) \in \mathbb{R}^n$. Wenn nun d ein Minimum in a annimmt, dann gelten notwendiger Weise die *Gauss'sche Normalengleichungen*

$$F^*F\,a \;=\; F^*\,y$$

Und in diesem Fall ist $y - f_a(x) \perp \mathrm{im}(F)$. Schließlich wenn $\mathrm{kn}(F) = 0$ ist (was $m \leq n$ voraussetzt), dann nimmt d ein absolutes Minimum an und F^*F ist invertierbar, so dass dieses Minimum auch berechnet werden kann, durch:

$$a \;=\; (F^*F)^{-1}\,F^*\,y$$

Weiterhin: ist die Funktion (zum Beispiel) f_1 konstant, dann stimmen der Durchschnitt der gefitteten Werte $f_a(x)$ und der Durchschnitt der gemessenen y Werte überein, d.h. es gilt

$$A(f_a(x)) \quad = \quad A(y)$$

Und damit besteht die folgende *Streuungszerlegung*: Die Varianz der gemessenen Werte $V(y)$ setzt sich zusammen aus der von x via f_a auf y übertragenen Varianz $V(f_a(x))$ und der Varianz $V(u)$ der Abweichung $u := y - f_a(x)$ der Messwerte vom Fit. D.h. dann gilt auch

$$V(y) \quad = \quad V(f_a(x)) + V(u)$$

Der Anteil der Streuung von y, der durch Übertragung (vermöge f_a) der Streuung von x erklärt werden kann wird *Bestimmtheitsmaß* $b(x \mid y)$ des Fits genannt. Aufgrund der Streuungszerlegung ist dies immer eine Zahl zwischen 0 und 1. Wir bezeichnen sie mit

$$b(x \mid y)^2 \quad := \quad \frac{V(f_a(x))}{V(y)}$$

Beweis:
Da die Funktion $\mathbb{R}^+ \to \mathbb{R}^+ : x \mapsto x^2$ streng monoton steigend ist, genügt es das Abstandsquadrat $q := d^2$ anstelle von d selbst zu minimieren. Wenn q also ein Minimum in a hat, dann ist a notwendiger Weise ein kritischer Punkt von q (vergleiche [Barner, Flohr, Analysis II], Kapitel 14.4). D.h. für alle $j \in 1 \ldots m$ muss gelten

$$\begin{aligned}
0 \quad &= \quad \partial_j q(a) \quad = \quad \sum_{k=1}^{n} \partial_j \left(\sum_{i=1}^{m} a_i f_i(x_k) - y_k \right)^2 \\
&= \quad \sum_{k=1}^{n} 2 f_j(x_k) \left(\sum_{i=1}^{m} a_i f_i(x_k) - y_k \right) \\
&= \quad 2 \sum_{i=1}^{m} a_i \sum_{k=1}^{n} f_i(x_k) f_j(x_k) - 2 \sum_{k=1}^{n} f_j(x_k) y_k \\
\Longrightarrow \quad \sum_{k=1}^{n} f_j(x_k) y_k \quad &= \quad \sum_{i=1}^{m} a_i \sum_{k=1}^{n} f_i(x_k) f_j(x_k)
\end{aligned}$$

Nach Definition der Matrix F gilt $f_{k,i} = f_i(x_k)$ und daher $f^*_{i,k} = f_i(x_k)$. Führen wir die Matrixmultiplikation aus, dann erkennt man, dass der (i,j)-te Koeffizient von F^*F sich folgendermaßen berechnet

$$[F^*F]_{i,j} \quad = \quad \sum_{k=1}^{n} f^*_{i,k} f_{k,j} \quad = \quad \sum_{k=1}^{n} f_i(x_k) f_j(x_k)$$

Da dieser Ausdruck symmetrisch in i und j ist, ist klar, dass F^*F eine symmetrische Matrix ist. Zudem kommt dieser Ausdruck in unseren partiellen Ableitungen vor, so dass wir die obige Gleichung umschreiben können, zu

$$[F^*\,y]_j \;=\; \sum_{k=1}^{n} f_j(x_k)y_k \;=\; \sum_{i=1}^{m} a_i\,[F^*F]_{i,j} \;=\; \sum_{i=1}^{m} [F^*F]_{j,i}\,a_i \;=\; [F^*F\,a]_j$$

Und da dies für alle $j \in 1\ldots m$ gilt, erhalten wir daraus die Gauß'schen Normalengleichungen $F^*F\,a = F^*\,y$ im Sinne von m-Tupeln.

Im Folgenden nehmen wir also an, es ist $\mathrm{kn}(F) = 0$. Zunächst sieht man leicht ein, dass damit F^*F ebenfalls injektiv sein muss: angenommen es wäre $F^*Fa = 0$, dann wäre auch

$$0 \;=\; \langle F^*Fa \mid a\rangle \;=\; \langle Fa \mid Fa\rangle \;= \|Fa\|^2$$

Wir hätten also auch $\|Fa\| = 0$ was wiederum $Fa = 0$ bedeutet. Nach Vorraussetzung für F folgt daraus aber $a = 0$. Damit ist F^*F injektiv und somit (als quadratische Matrix) bereits bijektiv. Mithin können die Gauß'schen Normalengleichungen eindeutig gelöst werden, durch

$$a \;=\; (F^*F)^{-1}\,F^*\,y$$

Aus den Gauß'schen Normalengleichungen folgt unmittelbar auch die Behauptung $y - f_a(x) \perp \mathrm{im}(F)$, denn ist $q = Fp$ ein beliebiger Bildvektor $p \in \mathbb{R}^m$, dann rechnet man nach, dass gilt

$$\langle q \mid y - f_a(x)\rangle \;=\; \langle Fp \mid y - Fa\rangle \;=\; \langle p \mid F^*y - F^*Fa\rangle \;=\; \langle p \mid 0\rangle \;=\; 0$$

Sei nun weiterhin f_1 eine konstante Funktion, etwa $f_1(x) = k$ für alle $x \in S$. Dann betrachten wir das Tupel $z := (1/k, 0, \ldots, 0) \in \mathbb{R}^m$. Offenbar gilt dafür $Fz = \mathbf{1} := (1, 1, \ldots, 1) \in \mathbb{R}^m$ und damit rechnet man dann nach, dass

$$\begin{aligned} A(y) - A(f_a(x)) \;&=\; A(y - f_a(x)) \;=\; A(y - Fa) \;=\; \frac{1}{n}\langle y - Fa \mid \mathbf{1}\rangle \\ &=\; \frac{1}{n}\langle y - Fa \mid Fz\rangle \;=\; \frac{1}{n}\langle F^*y - F^*Fa \mid z\rangle \;=\; 0 \end{aligned}$$

Insgesamt haben wir damit $A(y) = A(f_a(x))$ gezeigt. Als nächstes beweisen wir damit die Hilfsaussage $\sigma(y \mid Fa) = V(Fa)$ durch eine direkte Rechnung

$$\begin{aligned} \sigma(y \mid f_a(x)) \;&=\; A(y \cdot f_a(x)) - A(y)A(f_a(x)) \\ &=\; \frac{1}{n}\sum_{k=1}^{n} y_k f_a(x_k) - A(y)A(f_a(x)) \\ &=\; \frac{1}{n}\sum_{k=1}^{n} y_k \sum_{i=1}^{m} a_i f_i(x_k) - A(y)A(f_a(x)) \\ &=\; \frac{1}{n}\sum_{i=1}^{m} a_i \sum_{k=1}^{n} f_i(x_k)y_k - A(y)A(f_a(x)) \\ &=\; \frac{1}{n}\sum_{i=1}^{m} a_i\,[F^*y]_i - A(y)A(f_a(x)) \end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{n}\sum_{i=1}^{m} a_i \left[F^*Fa\right]_i - A(y)A(f_a(x)) \\
&= \frac{1}{n}\sum_{i=1}^{m} a_i \sum_{k=1}^{n} f_i(x_k)\left[Fa\right]_k - A(y)A(f_a(x)) \\
&= \frac{1}{n}\sum_{k=1}^{n} \left[Fa\right]_k \sum_{i=1}^{m} a_i f_i(x_k) - A(y)A(f_a(x)) \\
&= \frac{1}{n}\sum_{k=1}^{n} f_a(x_k)^2 - A(y)A(f_a(x)) \\
&= A(f_a(x)^2) - A(f_a(x))A(f_a(x)) \;=\; V(f_a(x))
\end{aligned}
$$

Und damit können wir dann schließlich die Aussage der Streuungszerlegung $V(y) = V(f_a(x)) - V(u)$ [wobei $u = y - f_a(x)$] beweisen. Wir rechnen einfach nach, dass (aufgrund der Linearität der Kovarianz in beiden Argumenten)

$$
\begin{aligned}
V(u) &= \sigma(u \mid u) \;=\; \sigma(y - f_a(x) \mid y - f_a(x)) \\
&= \sigma(y \mid y) - 2\sigma(y \mid f_a(x)) + \sigma(f_a(x) \mid f_a(x)) \\
&= V(y) - 2V(f_a(x)) + V(f_a(x)) \;=\; V(y) - V(f_a(x))
\end{aligned}
$$

□

Als erste Anwendung wollen wir uns folgendem Problem widmen: wir betrachten zwei Variablen x und y, zwischen denen ein Zusammenhang $y = f(x)$ vermutet wird, wobei $f \in \mathbb{R}[x]$ ein Polynom vom Grad m sein soll, d.h. f hat die Gestalt $f(x) = a_m x^m + \cdots + a_1 x + a_0$. Es werden n verschiedene Messungen gemacht, die die Wertepaare (x_1, y_1) bis (x_n, y_n) liefern. Die Frage ist nun, welche Abschätzung der Koeffizienten a_0, a_1 bis a_m von f die beste Übereinstimmung $y_i \approx f(x_i)$ liefert:

Satz 5.5: Polynomiale Regression:
Wir betrachten die Potenzen $f_i : \mathbb{R} \to \mathbb{R} : x \mapsto x^i$ (wobei $i \in 0 \ldots m$) und weiterhin sei $a = (a_0, a_1, \ldots, a_m) \in \mathbb{R}^{m+1}$ und wir bezeichnen

$$
F \;=\; \begin{pmatrix} 1 & x_1 & \cdots & x_1^m \\ \vdots & \vdots & & \vdots \\ 1 & x_n & \cdots & x_n^m \end{pmatrix}
$$

$$
f_a(x) \;=\; a_m x^m + \ldots + a_1 x + a_0 \;\in\; \mathbb{R}[x]
$$

Seien nun x_k und $y_k \in \mathbb{R}$ (wobei $k \in 1 \ldots n$) so gegeben, dass zumindest $m+1$ verschiedene x_k darunter sind, d.h. $m < \#\{\, x_k \mid k \in 1 \ldots n \,\}$. Dann ist die folgende $((m+1) \times (m+1))$-Matrix F^*F invertierbar und daher lässt sich die Gauss'sche Normalengleichung lösen

$$
F^*F \;=\; \left(\sum_{k=1}^{n} x_k^{i+j}\right)_{i,j \in 0 \ldots m}
$$

$$
a \;=\; \left(F^*F\right)^{-1} F^* y
$$

Satz 5.6: Lineare Regression:
Wir betrachten die Daten $x = (x_1, x_2, \ldots, x_n)$ und $y = (y_1, y_2, \ldots, y_n) \in \mathbb{R}^n$ so dass es mindestens 2 verschiedene x_k gibt, d.h. $2 \leq \#\{\, x_k \mid k \in 1 \ldots n \,\}$. Dann ist die *Regressionsgerade* $g(x) = ax + b$ die am nächsten durch die Punkte (x_k, y_k) verläuft - d.h. die den folgenden Abstand

$$d(a,b) \quad = \quad \sqrt{\sum_{k=1}^{n} (ax_k + b - y_k)^2}$$

minimiert - eindeutig bestimmt und ihre Koeffizienten a und b können mit Hilfe der folgenden Formeln berechnet werden:

$$\begin{aligned} a &= \frac{\sigma(x \mid y)}{\sigma(x)^2} \\ b &= A(y) - aA(x) \end{aligned}$$

Und wenn wir die Abweichung der Näherung von den Messwerten bezeichnen, mit $u_k := y_k - (ax_k + b)$ dann ist das Bestimmtheitsmaß der linearen Regression gerade der Korrelationskoeffizient, d.h. es gilt

$$b(x \mid y)^2 \quad = \quad \frac{V(ax+b)}{V(y)} \quad = \quad r(x \mid y)^2$$

$$\sigma(u) \quad = \quad \sqrt{1 - r(x \mid y)^2} \cdot \sigma(y) \quad = \quad \frac{1}{\sqrt{n}} d(a,b)$$

Bemerkung 5.7:
Der *Regressionskoeffizient* $r = r(x \mid y)$ ist ein Maß für den linearen Zusammenhang zwischen den x_k und den y_k. Für große Beträge von r (sagen wir für $|r| \in [0.6, 1]$) beobachtet man eine lineare Abhängigkeit. Das Vorzeichen von r gibt an ob der Zusammenhang gleichläufig ($r > 0$: wenn x steigt, dann auch y) oder gegenläufig ($r < 0$: wenn x steigt, dann sinkt y) ist. Bei der Interpretation muss man dennoch vorsichtig sein:

- Selbst wenn es einen klar erkennbaren Zusammenhang, etwa $r = 0.9$ gibt, muss es keinen inneren Grund dafür geben. Ein lehrreiches Beispiel dazu: die Korrelation zwischen der Geburtenrate (in 17 europäischen Ländern) und der Zahl der brütenden Storchenpaare beträgt $r = 0.62$ (und mit $p = 0.008$ ist dies sogar statistisch signifikant). Jedoch ist der Schluss, dass Störche die Babys liefern *nicht* zulässig. Dies ist lediglich der Tatsache geschuldet, dass es in hochzivilisierten Ländern sowohl weniger Babys als auch weniger Störche gibt [wir verweisen auf Robert Matthews, Storks deliver babies].

- Und selbst wenn kein linearer Zusammenhang besteht, etwa $r \approx 0$, dann heißt dies *nicht*, dass gar kein Zusammenhang zwischen den Daten besteht. Betrachtet man beispielsweise die sieben Datenpaare (k, k^2) für $k \in -3 \ldots 3$, so hängen diese offensichtlich funktional zusammen, unter $f(x) = x^2$. Nur der *lineare* Zusammenhang der Daten besteht nicht - es ist $r = 0$. Um den Zusammenhang aufzuspüren müsste man eine polynomiale Regression vom Grad ≥ 2 durchführen.

Anschaulich gesprochen zeigt die Zerlegung der Varianz, dass die Varianz von y in zwei Teile zerlegt werden kann: (1) aufgrund der linearen Beziehung $y \approx ax + b$ wird die Streuung der x auf y übertragen. Dies ist der *erklärbare* Anteil der Varianz von y. (2) die Varianz der Abweichung u, diese ist originär und kann nicht weiter erklärt werden. Und da $r = \sigma(ax+b)/\sigma(y)$ der Anteil der erklärten an der gesamten Streuung ist, liefert r also eine Einschätzung für die Güte der Annäherung $y \approx ax + b$.

Betrachten wir den Datensatz (x_k, y_k) so ist es möglich eine Approximation der Form $y \approx a_x x + b_x$ durchzuführen, aber ebenso eine Approximation der Form $x \approx a_y y + b_y$. Diese Koeffizienten werden also berechnet, durch:

$$\begin{array}{rclrcl} a_x & = & \sigma(x \mid y)/\sigma(x)^2 & a_y & = & \sigma(x \mid y)/\sigma(y)^2 \\ b_x & = & A(y) - a_x A(x) & b_y & = & A(x) - a_y A(y) \end{array}$$

Aufgrund der Symmetrie von $\sigma(x \mid y)$ ist damit aber klar, dass $r^2 = a_x a_y$. Will man nun $y = a_x x + b_y$ und $x = a_y y + b_y$ in das selbe Koordinatensystem einzeichnen, so müssen wir die zweite Gleichung umformen, zu $y = x/a_y - b_y/a_y$. Die beiden Linien stimmen genau dann überein, wenn $r = \pm 1$ und sie sind genau dann senkrecht, wenn $r = 0$ ist. Im allgemeinen kontrolliert r den Winkel α unter dem sich die beiden Regressionsgeraden schneiden:

$$\tan(\alpha) = \frac{a_x a_y - 1}{a_x + a_y} = \frac{r^2 - 1}{r^2} \cdot \frac{\sigma(x \mid y)}{\sigma(x)^2 + \sigma(y)^2}$$

Kapitel 6

Multivariate Statistik

In diesem Kapitel gehen wir auf ein paar Verallgemeinerungen der vorangegangenen Methoden ein, die über die bisherigen Grundkenntnisse hinaus gehen, die aber nicht fehlen dürfen, wenn man ernsthaft statistische Untersuchungen durchführen möchte. Im Einzelnen wollen wir auf höherdimensionale Probleme eingehen und für diese die Methoden der Kovarianzmatrix, der linearen Regression im n-dimensionalen Fall und die Faktorenanalyse erklären. Weitere Details dazu und die Beweise der hier gegebenen Sätze finden sich in meinem Online-Buch [Zeidler, Algebraic Statistics, auf www.wuala.com/zeidlerweb], das aber noch in den Kinderschuhen steckt.

Definition 6.1:
Wir betrachten zwei beliebige $(m \times n)$-Matrizen $(1 \leq m,\, n \in \mathbb{N})$ $X = (x_{i,j})$ und $Y = (y_{i,j}) \in \mathrm{mat}_{m,n}(\mathbb{R})$. Es bezeichne $\mathrm{row}_i(X)$ (mit $i \in 1 \ldots m$) die i-te Zeile und $\mathrm{col}_j(X)$ (mit $j \in 1 \ldots n$) die j-te Spalte von X. Dann definieren wir die **arithmetischen Mittel** von X als die folgenden Tupel:

$$\begin{aligned} A(X) &:= (A(\mathrm{col}_1(X)), \ldots, A(\mathrm{col}_n(X))) \in \mathbb{R}^n \\ A^*(X) &:= (A(\mathrm{row}_1(X)), \ldots, A(\mathrm{row}_m(X))) \in \mathbb{R}^m \end{aligned}$$

Zur Abkürzung sei nun $x_i := \mathrm{col}_i(X)$ und $y_j := \mathrm{col}_j(Y)$ (mit $i,\, j \in 1 \ldots n$), dann definieren wir ferner die **Kovarianz-Matrix** $\Sigma(X \mid Y) \in \mathrm{mat}_n(\mathbb{R})$ von X und Y und die **Varianz-Matrix** $V(X) = \Sigma(X \mid X) \in \mathrm{mat}_n(\mathbb{R})$ von X als die folgenden quadratischen Matrizen:

$$\begin{aligned} \Sigma(X \mid Y) &:= \begin{pmatrix} \sigma(x_1 \mid y_1) & \cdots & \sigma(x_1 \mid y_n) \\ \vdots & \ddots & \vdots \\ \sigma(x_n \mid y_1) & \cdots & \sigma(x_n \mid y_n) \end{pmatrix} \\ V(X) &:= \begin{pmatrix} \sigma(x_1 \mid x_1) & \cdots & \sigma(x_1 \mid x_n) \\ \vdots & \ddots & \vdots \\ \sigma(x_n \mid x_1) & \cdots & \sigma(x_n \mid x_n) \end{pmatrix} \end{aligned}$$

Beispiel 6.2:
Wir betrachten eine Klasse aus 30 Schülern, die in 10 verschiedenen Fächern unterrichtet werden (z.B. *Deutsch*, *Englisch*, *Mathematik* und so weiter). Dann wäre die Matrix X, die die Noten (von 1 bis 6) jedes Schülers in allen Fächern enthält, eine (30×10)-Matrix: die i-te Zeile bestünde aus den Noten von Schüler Nummer i, die j-te Spalte wären die Noten aller Schüler, die in Fach Nummer j vergeben wurden.

$$X \quad = \quad \begin{pmatrix} \cdots & \text{Noten von Schüler 1} & \cdots \\ & \vdots & \\ \cdots & \text{Noten von Schüler 30} & \cdots \end{pmatrix}$$

Dann wäre $A(X)$ das 10-Tupel der Notendurchschnitte der Klasse in den verschiedenen Fächern und $A^*(X)$ wäre das 30-Tupel der Notendurchschnitte der verschiedenen Schüler der Klasse.

Bemerkung 6.3:
Man sagt die Datenmatrix $X = (x_1, \ldots, x_n) \in (\mathbb{R}^m)^n = \mathrm{mat}_{m,n}(\mathbb{R})$ ist **zentriert**, wenn $A(X) = 0 \in \mathbb{R}^n$ gilt. Da $A(X) = (A(x_1), \ldots, A(x_n))$ ist, bedeutet das ausgeschrieben gerade $A(x_j) = 0$ für alle $j \in 1 \ldots n$. Man kann dies offenbar erreichen, durch $\widehat{X} := (\widehat{x}_1, \ldots, \widehat{x}_n)$ mit $\widehat{x}_j = x_j - A(x_j) \in \mathbb{R}^m$. Dann ist $\widehat{X}$ zentriert und in anderen Worten lautet $\widehat{X}$ gerade:

$$\widehat{X} \quad = \quad \begin{pmatrix} \vdots & & \vdots \\ x_1 & \cdots & x_n \\ \vdots & & \vdots \end{pmatrix} \quad - \quad \begin{pmatrix} \vdots & & \vdots \\ A(x_1) & \cdots & A(x_n) \\ \vdots & & \vdots \end{pmatrix}$$

Und da weiterhin $\sigma(x_i \mid x_j) = (\widehat{x}_i \bullet \widehat{x}_j)/m$ gilt, erhalten wir trivialer Weise (nach Definition der Varianzmatrix $V(X)$) die Identität der Matrizen

$$V(X) \quad = \quad \frac{1}{m} \widehat{X}^* \widehat{X}$$

Proposition 6.4:

(i) Wir betrachten beliebige $1 \leq m$, $n \in \mathbb{N}$ und zwei $(m \times n)$-Matrizen X und $Y \in \mathrm{mat}_{m,n}(\mathbb{R})$. Dann gilt offenbar $A(X^*) = A^*(X) \in \mathbb{R}^m$ und für die $(n \times n)$-Matrix $\Sigma(X \mid Y)$ gilt weiterhin

$$\Sigma(X \mid Y) \quad = \quad \Sigma(Y \mid X)^*$$

(ii) Seien weiterhin $a \in \mathbb{R}^n$ und $b \in \mathbb{R}$, dann betrachten wir die folgende Linearform $L : \mathbb{R}^n \to \mathbb{R} : x \mapsto a \bullet x + b$. Ist nun wieder $X \in \mathrm{mat}_{m,n}(\mathbb{R})$ eine reelle $(m \times n)$-Matrix, dann bezeichnen wir

$$L(X) \quad := \quad (L(\mathrm{row}_1 X), \ldots, L(\mathrm{row}_m X)) \quad \in \quad \mathbb{R}^m$$

Und damit ist das arithmetische Mittel der transformierten Daten $L(X)$ gerade wieder das mit L transformierte arithmetische Mittel $A(X)$ der Ausgangsdaten, formal geschrieben

$$A(L(X)) \quad = \quad L(A(X))$$

(iii) Sei nun $x \in \mathbb{R}^n$ und sei $T \in \mathrm{mat}_{m,n}(\mathbb{R})$ eine $(m \times n)$-Matrix. Dann gilt die folgende Gleichung für das arithmetische Mittel von Tx

$$A(Tx) \quad = \quad A(T) \bullet x$$

Und ist weiterhin $y \in \mathbb{R}^m$ dann erhalten wir auch die folgende Formel für die Kovarianz:

$$\sigma(Tx \mid y) \quad = \quad \frac{1}{m}(Tx \bullet y) - A(y) \cdot (A(T) \bullet x)$$

(iv) Wir betrachten nun $1 \leq m, n, p \in \mathbb{N}$ und die Matrizen $X \in \mathrm{mat}_{m,n}(\mathbb{R})$ und $T \in \mathrm{mat}_{n,p}(\mathbb{R})$. Dann lässt sich das arithmetische Mittel von XT wie folgt berechnen:

$$A(XT) \quad = \quad T^* A(X)$$

(v) Sei wieder $1 \leq n \in \mathbb{N}$, wir betrachten zwei zentrierte $(n \times n)$-Matrizen $X, Y \in \mathrm{mat}_n(\mathbb{R})$ d.h. es ist $A(X) = 0 = A(Y)$. Sei $T \in \mathrm{mat}_n(\mathbb{R})$ eine weitere $(n \times n)$-Matrix, dann gilt

$$\Sigma(TX \mid Y) \quad = \quad \Sigma(X \mid T^* Y)$$

Insbesondere: Ist T unitär (d.h. es ist $T^{-1} = T^*$), dann ist die Varianz invariant unter der Transformation mit T:

$$V(TX) \quad = \quad V(X)$$

Vorsicht: Es ist verführerisch zu glauben, die Identität $\Sigma(XT \mid Y) = \Sigma(X \mid YT^*)$ würde ebenfalls gelten. Diese gilt im allgemeinen aber nicht, sondern stimmt sogar nur in trivialen Fällen.

Im vorangegangenen Kapitel haben wir den funktionalen Zusammenhang $y = f(x)$ zweier reeller Variablen x und $y \in \mathbb{R}$ betrachtet. Es schließt sich die Frage an, was passiert, wenn f von mehreren Variablen $y = f(x_1, \ldots, x_n)$ abhängt. Wir betrachten hier aber nur den einfachsten Fall eines linearen Zusammenhangs, d.h. wir versuchen die Koeffizienten $a_1, \ldots, a_n$ und $b \in \mathbb{R}$

$$y \quad = \quad a_1 x_1 + a_2 x_2 + \ldots + a_n x_n + b$$

so zu finden, dass sie möglichst gut zu m Messwerten $(x_i, y_i) \in \mathbb{R}^n \times \mathbb{R}$ passen. Dabei bezeichnen wir $x_i = (x_{i,1}, \ldots, x_{i,n})$, d.h. x_i wird in die i-te Zeile der Datenmatrix X eingetragen. Wie man die beste Schätzung der Koeffizienten a_i und b erhält leistet der folgende

Satz 6.5:
Wir betrachten die Daten $(x_i, y_i) \in \mathbb{R}^n \times \mathbb{R}$ wobei $i \in 1 \dots m$ und $x_i = (x_{i,1}, \dots, x_{i,n}) \in \mathbb{R}^n$ sei. Damit definieren wir die $m \times (n+1)$ Matrix

$$A \;:=\; \begin{pmatrix} 1 & \cdots & x_1 & \cdots \\ \vdots & & \vdots & \\ 1 & \cdots & x_m & \cdots \end{pmatrix} \;=\; \begin{pmatrix} 1 & x_{1,1} & \cdots & x_{1,n} \\ \vdots & \vdots & & \vdots \\ 1 & x_{m,1} & \cdots & x_{m,n} \end{pmatrix}$$

Mit Hilfe dieser Matrix A können wir die m verschiedenen Gleichungen $a_1 x_{i,1} + \ldots + a_n x_{i,n} + b = y_i$ kompakt schreiben als $A(b,a) = y$, wobei natürlich $a = (a_1, \dots, a_n)$ und $y = (y_1, \dots, y_m)$ bezeichnet. Wenn nun $(b,a) \in \mathbb{R} \times \mathbb{R}^n$ den folgenden Abstand minimieren

$$d(a,b) \;:=\; \sqrt{\sum_{i=1}^{m} (a_1 x_{i,1} + \ldots + a_n x_{i,n} + b - y_i)^2} \quad \text{ist minimal}$$

dann erfüllt (b,a) notwendiger Weise die Gleichung $A^*A(b,a) = A^*y$. Insbesondere wenn die $(n+1) \times (n+1)$ Matrix A^*A invertierbar ist (z.B. wenn $\mathrm{kn}(A) = 0$), dann können wir die Koeffizienten $a \in \mathbb{R}^n$ und $b \in \mathbb{R}$ minimalen Abstands berechnen, durch

$$\begin{pmatrix} b \\ a \end{pmatrix} \;=\; (A^*A)^{-1} A^* y$$

Bemerkung 6.6:
Schreibt man das Gleichungssystem zu $A^*A(b,a) = A^*y$ explizit aus, dann nimmt es die folgende Form an: Seien wieder $(x_i, y_i) = (x_{i,1}, \dots, x_{i,n}, y_i) \in \mathbb{R}^{n+1}$ wobei $i \in 1 \dots m$ die empirischen Daten. Dann bezeichnen wir weiterhin $y = (y_1, \dots, y_m)$ und $X_j = (x_{1,j}, \dots, x_{m,j}) \in \mathbb{R}^m$ für $j \in 1 \dots n$. Das heißt X_j ist die $(j+1)$-te Spalte von A. Dann ist folgendes Gleichungssystem zu lösen:

$$\begin{pmatrix} X_1 \bullet X_1 & \cdots & X_1 \bullet X_n & X_1 \bullet \mathbf{1} \\ \vdots & & \vdots & \vdots \\ X_n \bullet X_1 & \cdots & X_n \bullet X_n & X_n \bullet \mathbf{1} \\ \mathbf{1} \bullet X_1 & \cdots & \mathbf{1} \bullet X_n & m \end{pmatrix} \begin{pmatrix} a_1 \\ \vdots \\ a_n \\ b \end{pmatrix} \;=\; \begin{pmatrix} X_1 \bullet y \\ \vdots \\ X_n \bullet y \\ \mathbf{1} \bullet y \end{pmatrix}$$

Proposition 6.7: Faktorenanalyse:
Wir betrachten eine zentrierte $(m \times n)$-Matrix $X \in \mathrm{mat}_{m,n}(\mathbb{R})$, das heißt X hat die Form $X = (x_1, \dots, x_n) \in (\mathbb{R}^m)^n$ wobei $A(x_j) = 0$ für alle $j \in 1 \dots n$. Sei weiterhin Σ die $(n \times n)$-Varianz-Matrix von X, das heißt $\Sigma := V(X) = (1/m) X^* X \in \mathrm{mat}_n(\mathbb{R})$. Dann gelten die folgenden Aussagen:

(i) Ist $a = (a_1, \dots, a_n) \in \mathbb{R}^n$ irgend ein n-Tupel, dann lässt sich die Varianz der linear transformierten Werte $Xa = a_1 x_1 + \cdots + a_n x_n \in \mathbb{R}^m$ berechnen vermöge der quadratischen Form

$$V(Xa) \;=\; \Sigma a \bullet a$$

(ii) Wir bezeichnen die Einheitssphäre im $\mathbb{R}^n$ mit $S^n := \{\, a \in \mathbb{R}^n \mid \|a\| = 1 \,\}$ und betrachten die Varianz-Funktion $v : S^n \to \mathbb{R} : a \mapsto V(Xa)$. Hat v in $s \in S^n$ ein lokales Extremum (d.h. es gibt ein $\varepsilon > 0$ so dass für alle $a \in S^n$ mit $\|a - s\| < \varepsilon$ gilt $v(s) \leq v(a)$ für ein lokales Minimum bzw. $v(a) \leq v(s)$ für ein lokales Maximum), dann ist s notwendiger Weise ein Eigenvektor von Σ.

(iii) Da Σ symmetrisch ist, gibt es eine Orthonormalbasis $\{\, t_1, \ldots, t_n \,\} \subseteq \mathbb{R}^n$ von $\mathbb{R}^n$ bestehend aus Eigenvektoren t_i von Σ. Wir bezeichnen die zugehörigen Eigenwerte mit $\lambda_i \in \mathbb{R}$, (d.h. es ist $\Sigma t_i = \lambda_i t_i$) und die Transformationsmatrix mit $T := (t_1, \ldots, t_n) \in \mathrm{mat}_n(\mathbb{R})$. Dann ist T offenbar unitär (d.h. es ist $T^{-1} = T^*$) und für jedes $a \in \mathbb{R}^n$ gilt

$$V(Xa) \;=\; \sum_{i=1}^{n} \lambda_i a_i^2$$

Insbesondere gilt $0 \leq V(Xt_i) = \lambda_i$ und damit ist klar, dass S^n unter v in ein Ellipsoid mit den Hauptachsen t_i verformt wird:

$$\{\, V(Xa) \mid a \in S^n \,\} \;=\; \left\{ \sum_{i=1}^{n} \lambda_i a_i^2 \;\middle|\; a \in S^n \right\}$$

Definition 6.8:
Wir betrachten eine beliebige $(m \times n)$-Matrix $X \in \mathrm{mat}_{m,n}(\mathbb{R})$ und bezeichnen ihre Zentralisierung mit $\widehat{X}$. Sei ferner Σ die $(n \times n)$-Varianzmatrix von X

$$\Sigma \;=\; \frac{1}{m} \widehat{X}^* \widehat{X}$$

Nun wählen wir eine Orthonormalbasis $\{\, t_1, \ldots, t_n \,\} \subseteq \mathbb{R}^n$ von $\mathbb{R}^n$, die aus Eigenvektoren von Σ besteht, und bezeichnen die zugehörigen Eigenwerte mit $\lambda_1, \ldots, \lambda_n \in \mathbb{R}$ (d.h. $\Sigma t_i = \lambda_i t_i$). Schließlich ordnen wir die Eigenwerte absteigend der Größe nach:

$$\lambda_1 \;\geq\; \lambda_2 \;\geq\; \ldots \;\geq\; \lambda_n$$

Ist nun $k \in 1 \ldots n$ eine Zahl, dann definieren wir die k-te **Hauptkomponente** von X als die folgende Linearkombination der $x_{i,j}$

$$\widehat{X}\, t_k \;\in\; \mathbb{R}^m$$

Beispiel 6.9:
Zur Illustration geben wir ein (rein künstliches) Beispiel, bestehend aus zwei Eigenschaften, die bei 4 Individuen ausgewertet werden. Sagen wir, x_1 liefert die Werte $x_1 = (3, 3, 4, 2)$ und $x_2 = (6, 7, 5, 6)$. Dann beginnen wir also mit der Datenmatrix

$$X \;=\; \begin{pmatrix} 3 & 6 \\ 3 & 7 \\ 4 & 5 \\ 2 & 6 \end{pmatrix}$$

Zunächst wird die Datenmatrix zentriert: die arithmetischen Mittel betragen: $A(x_1) = 3$ und $A(x_2) = 6$. Indem wir diese von X abziehen erhalten wir

$$\widehat{X} = \begin{pmatrix} 0 & 0 \\ 0 & 1 \\ 1 & -1 \\ -1 & 0 \end{pmatrix}$$

Mit Hilfe dieser zentrierten Matrix können wir die Varianzmatrix von X berechnen: $\Sigma = V(X) = (1/4)\widehat{X}^*\widehat{X}$ und dies führt auf

$$\Sigma = \frac{1}{4}\begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}$$

Das charakteristische Polynom einer (2×2)-Matrix ist bekanntlich $c(\lambda) = \lambda^2 - \mathrm{trace}(\Sigma)\lambda + \det(\Sigma)$ und dieses hat die Nullstellen $\lambda_1 = 3/4$ und $\lambda_2 = 1/4$. Die zugehörigen Eigenvektoren sind $e_1 = (1, -1)$ und $e_2 = (1, 1) \in \mathbb{R}^2$. Indem wir diese Vektoren normieren finden wir die Transformationsmatrix

$$T = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}$$

Also was bedeuten diese Eigenvektoren nun? Der Erste e_1 sagt uns, dass die Wertepaare (x_1, x_2) im wesentlichen durch den Wert $x_1 - x_2$ bestimmt werden. Dieser Wert trägt $3/4$ zur Varianz bei. Umgekehrt sagt uns der Zweite e_2, dass der Wert $e_1 + e_2$ relativ konstant ist - er trägt nur $1/4$ zur Varianz bei. Daher ist $x_1 + x_2$ recht uninteressant und es ist der Wert $x_1 - x_2$ an dem wir die 4 Individuen besser unterscheiden können.

$$\widehat{X}t_1 = \frac{1}{\sqrt{2}}\begin{pmatrix} 0 \\ -1 \\ 2 \\ -1 \end{pmatrix} \quad \text{und} \quad \widehat{X}t_2 = \frac{1}{\sqrt{2}}\begin{pmatrix} 0 \\ 1 \\ 0 \\ -1 \end{pmatrix}$$

Kapitel 7

Laplace Räume

Beispiel 7.1:
Wir betrachten einen (idealen, sechsseitigen) Würfel. D.h. der Würfel hat für jede Zahl $s \in 1 \ldots 6$ die gleiche Wahrscheinlichkeit von $1/6$ diese Zahl zu würfeln. Wir betrachten nun aber zwei Würfel, die zusammen geworfen werden. Die Würfel seien unabhängig voneinander (also nicht verklebt oder magnetisch oder so etwas). Dann hat jede Kombination von Würfelergebnissen $(s,t) \in (1 \ldots 6)^2$ die gleiche Wahrscheinlichkeit geworfen zu werden. Und da es 36 solcher *Elementarereignisse* gibt, beträgt die Wahrscheinlichkeit also konstant $1/36$. Wir interessieren uns nun aber nicht für die einzelnen, speziellen Würfelkombinationen (s,t), sondern für die Summe $s+t$ der beiden Augenzahlen. Dann finden wir folgende Tabelle:

$s+t$	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

An dieser Tabelle sieht man, dass genau die Kombinationen $(1,3)$, $(2,2)$ und $(3,1)$ die Summe 4 ergeben. Es besteht also die folgende Gleichheit

$$\left\{ (s,t) \in (1 \ldots 6)^2 \mid s+t=4 \right\} \quad = \quad \{ (1,3),(2,2),(3,1) \}$$

Will man wissen wie hoch die Wahrscheinlichkeit für das *Ereignis* ist, mit zwei Würfeln die Augensumme 4 zu erreichen, muss man also die Einzelwahrscheinlichkeiten der drei Elementarereignisse $(1,3)$, $(2,2)$ und $(3,1)$ addieren

$$p(s+t=4) \quad = \quad p(1,2)+p(2,2)+p(2,1) \quad = \quad \frac{3}{36} \quad = \quad \frac{1}{12}$$

Wir müssen also folgende Begriffe formalisieren: Das Ergebnis unseres Zufallsexperimentes (Werfen zweier Würfel) liegt in irgend einer *Ereignismenge* S. Jedes *Elementarereignis* $s \in S$ hat eine gewisse *Wahrscheinlichkeit* $p(s)$ einzutreten. Auf dem Ereignisraum S lebt eine *Zufallsvariable* $X : S \to \mathbb{R}$. Diese ordnet jedem möglichen Ereignis ein Ergebnis zu, in unserem Beispiel war $X(s,t) = s+t$.

Die Wahrscheinlichkeit für das *Ereignis*, dass $X = a \in \mathbb{R}$ wird, erhält man als Summe der Einzelwahrscheinlichkeiten $p(s)$ aller Elementarereignisse $s \in S$, für die $X(s) = a$ gilt:

$$p(X = a) \quad = \sum_{X(s)=a} p(s)$$

Eine *Stichprobe* $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$ von X ist also die Liste der Ergebnisse x_k bei n-maliger Durchführung des Zufallsexperimentes. Sie liegt also in der Bildmenge der Zufallsvariablen. Ein zentrales Anliegen der Statistik ist es die zugrunde liegende Wahrscheinlichkeit aus einer Stichprobe zu schätzen.

Definition 7.2:
Wir nennen das Paar (S, p) einen **Laplace Raum** (oder ausführlicher einen *Laplace'schen Wahrscheinlichkeitsraum*) wenn $S \neq \emptyset$ eine nicht-leere Menge ist und $p : S \to \mathbb{R}$ eine summierbare Zahlenfamilie auf S ist, für die gilt

(1) $\forall\, s \in S \;:\; p(s) \geq 0$

(2) $\sum_{s \in S} p(s) \;=\; 1$

Und ist $A \subseteq S$ eine beliebige Teilmenge von S, dann definieren wir die Schreibweise (die Summe konvergiert, wegen der Summierbarkeit von p)

$$p(A) \;:=\; \sum_{s \in A} p(s) \;\in\; [0, 1]$$

Unter einer (reellen) **Zufallsvariable** auf (S, p) verstehen wir eine beliebige Abbildung der Form $X : S \to \mathbb{R}$. Und ist $a \in \mathbb{R}$ eine Zahl, dann definieren wir die Notationen

$$\begin{aligned} p(X = a) &:= p(\{\, s \in S \mid X(s) = a \,\}) \\ p(X < a) &:= p(\{\, s \in S \mid X(s) < a \,\}) \\ p(X > a) &:= p(\{\, s \in S \mid X(s) > a \,\}) \end{aligned}$$

Bemerkung 7.3:

- Der Begriff der *Summierbarkeit* wurde in (2.5) eingeführt. Wer damit noch hadert, der kann sich ohne weiteres auf *endliche* Mengen S beschränken. In diesem Fall sind alle Summen endlich und man braucht keine Konvergenz zu betrachten.

- Aufgrund der Summierbarkeit muss die Menge $T := \{\, s \in S \mid p(s) > 0 \,\}$ der möglichen Ereignisse (für $s \notin T$ ist $p(s) = 0$) *abzählbar* sein. D.h. es gibt eine surjektive Abbildung $\mathbb{N} \twoheadrightarrow T$, die die Abzählung der Elemente von T ermöglicht. Daher ist es mit Laplace-Räumen nicht möglich Probleme zu modellieren, bei denen überabzählbar viele mögliche Ergebnisse vorkommen können.

- Man beachte, dass wir durch die eben eingeführte Notation von der Funktion $p : S \to \mathbb{R}$ zu einer Funktion $p : \mathcal{P}(S) \to \mathbb{R}$ übergegangen sind. Dabei ist $\mathcal{P}(S)$ die *Potenzmenge* von S, d.h. die Menge aller Teilmengen $\mathcal{P}(S) = \{ A \mid A \subseteq S \}$ von S. Wir werden sehen, dass es unmöglich ist *jeder* Teilmenge von $\mathbb{R}$ ein vernünftiges Volumen zuzuordnen. Das bedeutet, dass wir für überabzählbare S nicht von jeder Teilmenge $A \subseteq S$ eine Wahrscheinlichkeit $p(A)$ angeben können. In späteren Kapiteln widmen wir uns dem Problem, wie man sich auf geeignete Mengen zurückziehen kann.

- Pierre Simon Laplace (1749–1827) selbst hatte folgende Einsicht: haben unter n Elementarereignissen $S = \{ s_1, \ldots, s_n \}$ alle dieselbe Wahrscheinlichkeit einzutreten, dann ist die Wahrscheinlichkeit das Ereignis $A \subseteq S$ zu erhalten gegeben, durch:

$$p(A) \quad = \quad \frac{\#A}{n} \quad = \quad \frac{\text{Zahl günstiger Fälle}}{\text{Zahl möglicher Fälle}}$$

- Zum Begriff einer Zufallsvariablen ist folgendes zu bemerken: Führt man ein Zufallsexperiment aus (z.B: das Ziehen von n Losen aus einer Lostrommel), so erhält man ein n-Tupel $(s_1, \ldots, s_n) \in S^n$ von Zufallsergebnissen, etwa den gezogenen Losnummern. Jedem Ergebnis $s_i \in S$ wird dann ein Wert $x_i \in \mathbb{R}$ zugeordnet - etwa die Gewinnsumme des Loses. Es sind natürlich auch andere Wertebereiche als $\mathbb{R}$ denkbar, aber das ist jetzt unerheblich. *A posteriori* - also nach dem Experiment - ist das Ergebnis also ein n-Tupel $x = (x_1, \ldots, x_n) \in \mathbb{R}$ reeller Zahlen.

  Die Frage ist, wie wir das *a priori* - also vor der Durchführung des Experimentes - modellieren können. Da ist ja noch jedes $s \in S$ möglich, wir kennen nur die Wahrscheinlichkeit $p(s)$, dass dabei s heraus kommt. Man betrachtet daher ganze Abbildungen $X : S \to \mathbb{R}$, die jedem *möglichen* Ergebnis $s \in S$ den daraus resultierenden Wert $X(s) \in \mathbb{R}$ zuordnen. Bleiben wir beim Beispiel der Lostrommel: haben wir die Gewinne $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$ gezogen, dann beträgt der durchschnittliche Gewinn einer Ziehung offenbar

$$A(x) \quad = \quad \frac{1}{n} \sum_{i=1}^{n} x_i$$

  Die Wahrscheinlichkeiten $p(s)$ sind dann bereits in der Ziehung x enthalten. Wie sieht es aber a priori aus? Die Wahrscheinlichkeit das Los $s \in S$ zu ziehen beträgt $p(s)$, der Gewinn in diesem Fall wäre $X(s)$. Also ist $X(s)p(s)$ der anteilige Gewinn des Loses s. Um eine Schätzung für die Gewinnsumme zu erhalten, müssen wir über all die möglichen Ausgänge $s \in S$ des Versuchs summieren. Diese Überlegung führt zum Begriff des *Erwartungswertes* $E(X)$ von X, auf den wir später noch genauer eingehen werden

$$E(X) \quad = \quad \sum_{s \in S} X(s)p(s)$$

Diese Abschätzung sagt natürlich nichts über den Einzelfall - die nächste Ziehung - aus. Erst wenn wir viele Ziehungen machen, wird sich $A(x)$ dem Wert $E(X)$ nähern. Und auch das nur, wenn wir immer denselben Versuch wiederholen - bei einer Ziehung von Losen ohne Zurücklegen ist das beispielsweise nicht der Fall. Ist der Hauptgewinn erst gezogen, kann er nicht nochmal gezogen werden, die Wahrscheinlichkeitsverteilung würde sich also in jedem Zug ändern.

Proposition 7.4: Kolmogorov-Eigenschaften:
Es sei (S, p) ein Laplace-Raum und A, $B \subseteq S$ seien zwei beliebige Teilmengen von S. Dann gelten die folgenden Aussagen:

(1) $p(\emptyset) = 0$

(2) $p(S) = 1$

(3) $p(A) \geq 0$

(4) $A \subseteq B \implies p(A) \leq p(B)$

(5) $p(A \cup B) = p(A) + p(B) - p(A \cap B)$

(6) $p(B \setminus A) = p(B) - p(A \cap B)$

Definition 7.5:
Ist (S, p) ein Laplace-Raum und sind A und $B \subseteq S$ zwei Teilmengen von S, wobei $p(B) \neq 0$, dann definieren wir die **bedingte Wahrscheinlichkeit** von A unter der Bedingung B, durch:

$$p(A \mid B) \quad = \quad \frac{p(A \cap B)}{p(B)}$$

Und die Ereignisse A, $B \subseteq S$ heißen **unabhängig**, falls die folgende Eigenschaft (c) erfüllt ist. Dazu äquivalent sind die Eigenschaften (a) (falls $p(B) \neq 0$) und (b) (falls $p(A) \neq 0$)

(a) $p(A \mid B) = p(A)$

(b) $p(B \mid A) = p(B)$

(c) $p(A \cap B) = p(A) \cdot p(B)$

Bemerkung 7.6:
Man beachte, dass für allgemeine Teilmengen A, $B \subseteq S$ gilt: $p(A \mid B) \cdot p(B) = p(A \cap B)$. D.h. die Wahrscheinlichkeit für A *und* B setzt sich zusammen aus der bedingten Wahrscheinlichkeit von A unter der Bedingung B und der Wahrscheinlichkeit für B selbst. Dies ist hilfreich, da es oft einsichtig ist, wie $p(B)$ und $p(A \mid B)$ lauten, aber $p(A \cap B)$ direkt ist nur schwer abzusehen. Die bedingte Wahrscheinlichkeit ist sozusagen die Hintertür zu 'A und B'. Sei nun $\{ S_1, \ldots, S_k \} \subseteq \mathcal{P}(S)$ eine endliche *Partition* von S, d.h. es gelte $S = S_1 \cup \ldots \cup S_k$ und für $i \neq j \in 1 \ldots k$ seien $S_i \cap S_j = \emptyset$ und $p(S_i) \neq 0$. Dann erhalten wir daraus den *Satz von der totalen Wahrscheinlichkeit*:

$$p(A) \quad = \quad \sum_{i=1}^{k} p(A \mid S_i) \cdot p(S_i)$$

Anschaulich gesprochen bedeutet das folgendes: Wir haben den Ereignisraum S in verschiedene Fälle S_1 bis S_k untergliedert. Die Fälle dürfen sich nicht überschneiden und müssen alle Möglichkeiten ausschöpfen. Dann können wir die absolute Wahrscheinlichkeit $p(A)$ für das Ereignis A zusammensetzen durch die bedingten Wahrscheinlichkeiten $p(A \mid S_i)$ in den einzelnen Fällen und die jeweilige Wahrscheinlichkeit $p(S_i)$ unter der S_i vorkommt.

Zum Beweis: Da die S_i eine Partition von S bilden, sind die $A \cap S_i$ eine Partition von A. Daher ist $p(A) = \sum_i p(A \cap S_i)$. In dieser Gleichung ersetzt man nur noch $p(A \cap S_i) = p(A \mid S_i) \cdot p(S_i)$ und schon ist man fertig.

Beispiel 7.7:

- Der Begriff der bedingten Wahrscheinlichkeit ist die Übertragung der bedingten Häufigkeit in die Wahrscheinlichkeitstheorie. Wir erinnern an unser Beispiel zu Beginn von Kapitel 5: Die bedingte Häufigkeit arbeitslos zu sein unter der Bedingung ein Studium abgeschlossen zu haben war: Anzahl der arbeitslosen Akademiker pro Anzahl aller Akademiker. Das Verhältnis ist dasselbe, wenn wir statt der Anzahl die Häufigkeit nehmen. Also die Häufigkeit der Kombination A und B pro Häufigkeit von B. Führt man genügend Stichproben durch, dann geht die Häufigkeit aber in die Wahrscheinlichkeit über. Und deswegen haben wir $p(A \mid B)$ gerade definiert, als p von A und B pro p von B.

- Ihr neuer Kollege hat zwei Kinder, eines davon ist ein Mädchen. Wie groß ist die Wahrscheinlichkeit dafür, dass das andere Kind ein Junge ist? Naiv würde man sofort 50% sagen. Eine genauere Betrachtung zeigt aber, das das nicht stimmt: Bei 2 Kindern gibt es 4 Möglichkeiten für deren Geschlecht: (w,w), (w,m), (m,w) und (m,m). Die Kombination (m,m) können wir im Vorfeld ausschließen, da wir ja bereits wissen, das eines davon ein Mädchen ist. Es bleiben also die 3 Möglichkeiten (w,w), (w,m) und (m,w). In 2 dieser Fälle ist aber ein Junge darunter, nur in einem Fall sind beides Mädchen. Da alle 4 Fälle gleich wahrscheinlich sind, erhält man: die Wahrscheinlichkeit, dass das andere Kind ein Junge ist, ist 2/3!

- Bei einer Spielshow wählt der Kandidat eines von 3 Toren A, B oder C. Hinter einem der Tore ist der Gewinn, zwei der Tore sind Nieten. Die Chance auf das richtige Tor zu tippen ist also 1/3. Nachdem der Kandidat gewählt hat, öffnet der Showmaster ein drittes Tor - weder das gewählte, noch das mit dem Gewinn - und fragt 'Wollen Sie jetzt doch das andere Tor nehmen?' Was würden Sie tun? Denken wir nach: wenn Sie zuerst richtig gewählt hatten (= Bedingung *richtig*), dann müssen Sie das richtige Tor jetzt verlassen. Ihre Gewinnchance (= Ereignis *Gewinn*) ist in diesem Fall also gleich 0:

$$p(\text{Gewinn} \mid \text{richtig}) \quad = \quad 0$$

Hatten Sie aber zuerst das falsche Tor gewählt (= Bedingung *falsch*), dann weichen Sie jetzt zwingend auf das richtige Tor aus (denn eine Niete verlassen Sie und eine Niete hat Ihnen der Showmaster gezeigt, es bleibt nur das Gewinntor). Ihre Gewinnchance ist in diesem Fall ist:

$$p(\text{Gewinn} \mid \text{falsch}) \quad = \quad 1$$

Die beiden Fälle *richtig* und *falsch* bilden offensichtlich eine Partition des Raumes S. Es gilt also der Satz der totalen Wahrscheinlichkeit:

$$\begin{aligned} p(\text{Gewinn}) \quad &= \quad \begin{array}{l} p(\text{Gewinn} \mid \text{richtig}) \cdot p(\text{richtig}) \\ +p(\text{Gewinn} \mid \text{falsch}) \cdot p(\text{falsch}) \end{array} \\ &= \quad 0 \cdot \frac{1}{3} + 1 \cdot \frac{2}{3} \quad = \quad \frac{2}{3} \end{aligned}$$

Wenn Sie sich umentscheiden steigt Ihre Gewinnchance also auf 2/3! Allgemeiner kann man diese Spiel mit n Toren spielen. Dann sind die bedingten Wahrscheinlichkeiten wieder $p(\text{Gewinn} \mid \text{richtig}) = 0$ und (weil 2 Nieten entfallen) $p(\text{Gewinn} \mid \text{falsch}) = 1/(n-2)$. Insgesamt steigt Ihre Gewinnchance beim Umentscheiden auf:

$$p(\text{Gewinn}) \quad = \quad 0 \cdot \frac{1}{n} + \frac{1}{n-2} \cdot \frac{n-1}{n} \quad = \quad \frac{n-1}{n-2} \cdot \frac{1}{n}$$

- Wir werden uns später noch sehr viel intensiver mit dem Begriff der Unabhängigkeit beschäftigen. Dass es dafür nicht reicht für drei Ereignisse A, B und $C \subseteq S$ zu fordern, dass $p(A \cap B \cap C) = p(A) \cdot p(B) \cdot p(C)$ ist wollen wir aber schon jetzt geben. Sei $S = 1 \ldots 8$ unter der Gleichverteilung $p(s) = 1/8$ für alle $s \in S$. Dann betrachten wir $A := \{1, 2, 4, 6\}$ und $B = C = \{1, 3, 5, 7\}$. Offenbar ist $p(A \cap B \cap C) = p(1) = 1/8 = (1/2)^3 = p(A) \cdot p(B) \cdot p(C)$. Jedoch sind B und C nicht unabhängig, da $p(B \cap C) = p(B) = 1/2$ und $p(B) \cdot p(C) = (1/2)^2 = 1/4$ sind.

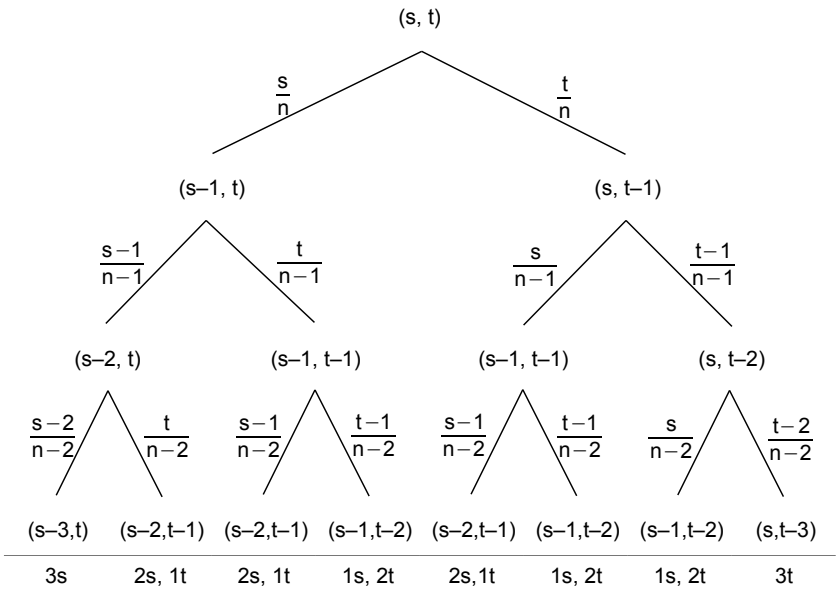
- Wir haben eine Urne mit s schwarzen und t türkisfarbenen Kugeln vorliegen und es sei $n = s + t$. Die Wahrscheinlichkeit bei 3 Ziehungen, *mit* Zurücklegen, erst eine schwarze, dann eine türkisfarbene und wieder eine schwarze Kugel zu ziehen beträgt

$$p(s)p(t)p(s) \quad = \quad \frac{s}{n} \cdot \frac{t}{n} \cdot \frac{s}{n} \quad = \quad \frac{s^2 t}{n^3}$$

Nun dasselbe *ohne* Zurücklegen - einmal gezogen, bleibt die Kugel draußen. Ziehen wir die erste schwarze Kugel, bleiben nur $s-1$ schwarze Kugeln zurück. Die Wahrscheinlichkeit nun eine türkisfarbene Kugel zu ziehen hat sich also geändert, zu $t/(n-1)$. Wurde auch diese Kugel gezogen sind also noch $s-1$ schwarze Kugeln in der Urne, die insgesamt nur noch $n-2$ Kugeln enthält. Die Wahrscheinlichkeit, dass die nächste gezogene Kugel schwarz ist, ist jetzt also $(s-1)/(n-2)$. Insgesamt beträgt die Wahrscheinlichkeit der Zugfolge also

$$\frac{s}{n} \cdot \frac{t}{n-1} \cdot \frac{s-1}{n-2}$$

Am übersichtlichsten wird die Ziehung in einem *Baumdiagramm*: hier steht an jedem Knoten der gegenwärtige Zustand (die verbliebenen Kugeln in der Urne). An jedem Ast steht die entsprechende Übergangswahrscheinlichkeit - die bedingte Wahrscheinlichkeit für den neuen Zustand unter der Bedingung des alten Zustands:



Bemerkung 7.8: Pfadregel:
Die bedingten Wahrscheinlichkeiten treten also als Übergangswahrscheinlichkeiten auf: Ist das System in Zustand B, dann ist $p(A \mid B)$ die Wahrscheinlichkeit dafür, dass im nächsten Schritt A heraus kommt. Multipliziert man all diese Wahrscheinlichkeiten entlang der Äste in einem solchen Baum von der Wurzel bis zu einem der Blätter herunter, dann erhält man die Gesamtwahrscheinlichkeit *entlang dieses Pfades* das Ergebnis des Blattes zu erreichen.

Mathematisch ausgedrückt ist dies die folgende Tatsache (die sich leicht durch Induktion über m beweisen lässt): Sind $A_1, A_2, \ldots, A_m \subseteq S$ irgend welche Ereignisse, mit $p(A_1 \cap A_2 \cap \ldots \cap A_{m-1}) \neq 0$, dann gilt:

$$p(A_1 \cap A_2 \cap \ldots \cap A_m) \;=\; p(A_1) \cdot \prod_{k=1}^{m-1} p(A_{k+1} \mid A_1 \cap \ldots \cap A_k)$$

Satz 7.9: Satz von Bayes:
Sei wieder (S, p) eine Laplace-Raum und die Teilmenge $A \subseteq S$ ein beliebiges Ereignis. Wir betrachten wie zuvor eine Partition $\{ S_1, S_2, \ldots, S_k \}$ von S. Dabei gelte $p(A) \neq 0$ und $p(S_i) \neq 0$ für alle $i \in 1 \ldots k$. Dann gilt für alle $j \in 1 \ldots k$ umgekehrt auch:

$$p(S_j \mid A) \;=\; \frac{p(A \mid S_j)\, p(S_j)}{p(A)} \;=\; \frac{p(A \mid S_j)\, p(S_j)}{\sum_{i=1}^{k} p(A \mid S_i)\, p(S_i)}$$

Beispiel 7.10:
Jetzt wird es leicht politisch: Wir untersuchen die Beweiskraft von Geständnissen. Wir bezeichnen die folgenden Ereignisse: $S :=$ der Angeklagte ist schuldig, $U :=$ der Angeklagte ist unschuldig und $G :=$ der Angeklagte hat ein Geständnis abgelegt. Offensichtlich bilden S und U eine Partition des Raumes und damit können wir die Wahrscheinlichkeit, das der Angeklagte schuldig ist, unter der Annahme, dass er gestanden hat, mit dem Satz von Bayes ausdrücken:

$$p(S \mid G) \quad = \quad \frac{p(G \mid S)p(S)}{p(G \mid U)p(U) + p(G \mid S)p(S)}$$

Es bezeichne weiterhin $s := p(S)$ und $r := p(G \mid U)/p(G \mid S)$. D.h. s ist die (unbekannte) Schuldwahrscheinlichkeit des Angeklagten und r gibt an um wieviel (un)wahrscheinlicher es ist, dass ein Unschuldiger gesteht, als ein Schuldiger. Wegen $p(U) = 1 - s$ erhalten wir (durch Kürzen von $p(G \mid S)$):

$$p(S \mid G) \quad = \quad \frac{s}{r(1-s) + s}$$

Das Gericht ist natürlich davon überzeugt, dass ein Geständnis dafür spricht, dass der Angeklagte tatsächlich schuldig ist. D.h. der Richter glaubt an $p(S) \leq p(S \mid G)$. Mit obiger Formel lautet dies also $s \leq s/(r(1-s)+s)$ und eine leichte Rechnung liefert daraus gerade $r \leq 1$, also $p(G \mid U) \leq p(G \mid S)$. Das heißt ein Geständnis erhöht die Schuldwahrscheinlichkeit nur, wenn eine schuldige Person eher gesteht, als eine unschuldige Person. Auf den ersten Blick denkt man: 'ja das ist doch wohl so'. Aber stimmt das überhaupt? Es kommt tatsächlich vor, dass ein Unschuldiger unter Druck Taten gesteht, die er gar nicht begangen hat. Und wenn wir an Terroristen denken - die wurden in Trainingslagern ausgebildet Druck auszuhalten, der Normalbürger aber nicht. Mit einem Blick auf Guantanamo sollte man also fest halten: Bei mutmaßlichen Terroristen *senkt* ein Geständnis, das unter Druck gegeben wurde, die Schuldwahrscheinlichkeit!

Bemerkung 7.11:
Während wir in den vorangegangenen Beispielen die Wahrscheinlichkeit $p(s)$ aus den bedingten Wahrscheinlichkeiten zusammensetzen konnten, ist es in vielen Fällen nur möglich die Wahrscheinlichkeit zu schätzen - oder empirisch zu untersuchen. In letzterem Fall ist es intuitiv klar, dass die relative Häufigkeit $h(s)$ für das Ereignis $s \in S$ in etwa die Wahrscheinlichkeit $p(s)$ des Zustands sein muss. Und die Übereinstimmung muss umso besser sein, je größer der Umfang der Stichprobe ist.

D.h führen wir eine Stichprobe $f : 1 \ldots n \to S$ vom Umfang n im Laplace-Raum (S, p) aus, dann nähern sich zu jedem Ereignis $s \in S$ die relativen Häufigkeiten $h(s) = \#\{\, i \in 1 \ldots n \mid f(i) = s \,\}$ mit wachsendem n der Wahrscheinlichkeit $p(s)$ an. Dies nennt man das *Gesetz der großen Zahlen*

$$\forall\, s \in S \quad : \quad h(s) \to p(s) \quad \text{für} \quad n \to \infty$$

Werden die verschiedenen Ereignisse $s \in S$ nun noch mit reellen Werten belegt $X : S \to \mathbb{R}$ dann haben wir in (4.3.(ii)) gesehen, wie sich der Durchschnitt der Werte $x := Xf : 1 \ldots n \to \mathbb{R} : i \mapsto X(f(i))$ berechnen lässt

$$A(x) \quad = \quad \sum_{s \in S} h(s)X(s)$$

Setzt man nun den Grenzübergang $h(s) \to p(s)$ an, dann muss also die Folge der arithmetischen Mittel $A(f(x))$ konvergieren, gegen den Grenzwert

$$A(x) \to \sum_{s \in S} p(s)X(s) \quad \text{für} \quad n \to \infty$$

Den Grenzwert der arithmetischen Mittel von $f(x)$ nennen wir den Erwartungswert von X. Analog bezeichnet man den Grenzwert der Varianz von $f(x)$ wiederum als Varianz von X:

Definition 7.12:
Sei (S, p) ein Laplace-Raum und $X\ Y : S \to \mathbb{R}$ zwei Zufallsvariable auf S. Dann definieren wir den **Erwartungswert** $E(X)$ und die **Varianz** $V(X)$ von X bzw. die **Kovarianz** $\sigma(X \mid Y)$ von X und Y durch

$$\begin{aligned}
E(X) \quad &:= \quad \sum_{s \in S} p(s)X(s) \\
\sigma(X \mid Y) \quad &:= \quad \sum_{s \in S} p(s)(X(s) - E(X)) \cdot (Y(s) - E(Y)) \\
&= \quad E(XY) - E(X) \cdot E(Y) \\
V(X) \quad &:= \quad \sigma(X \mid X) \quad = \quad \sum_{s \in S} p(s)(X(s) - E(X))^2
\end{aligned}$$

Beispiel 7.13: Chuck a Luck:
Beim Spiel *Chuck a Luck* zahlt der Spieler einen Einsatz e um auf eines von 6 nummerierten Feldern setzen zu können. Sagen wir das gewählte Feld sei $f \in 1 \ldots 6$. Nun wirft er 3 (sechsseitige) Würfel. Und damit gewinnt er so viele Euro, wie er Würfel hat, die die Augenzahl f zeigen. Die Frage ist: bei welchem Einsatz e ist das Spiel fair? Dazu müssen wir zunächst das Spiel modellieren:

$$S \quad := \quad (1 \ldots 6)^3 \quad = \quad \{\, (d_1, d_2, d_3) \mid d_i \in 1 \ldots 6 \,\}$$

Die Zufallsfunktion, die jedem Elementarereignis $s = (d_1, d_2, d_3) \in S$ die Zahl der Würfel zuordnet, die f zeigen bezeichnen wir mit K, formal also

$$K \ : \ S \to \mathbb{N} \ : \ (d_1, d_2, d_3) \mapsto \#\,\{\, i \in 1 \ldots 3 \mid d_i = f \,\}$$

Diese Zahl folgt offensichtlich einer 3-stufigen Binomialverteilung mit der Wahrscheinlichkeit von 1/6 die Zahl f zu würfeln und 5/6 eine andere Zahl zu würfeln. Damit ergeben sich die folgenden Wahrscheinlichkeiten

$$\begin{aligned}
p(K=0) &= \left(\tfrac{5}{6}\right)^3 &&= \frac{125}{216}\\
p(K=1) &= 3\cdot\tfrac{1}{6}\cdot\left(\tfrac{5}{6}\right)^2 &&= \frac{75}{216}\\
p(K=2) &= 3\cdot\left(\tfrac{1}{6}\right)^2\cdot\tfrac{5}{6} &&= \frac{15}{216}\\
p(K=3) &= \left(\tfrac{1}{6}\right)^3 &&= \frac{1}{216}
\end{aligned}$$

Die Zufallsvariable für den Gewinn ist offenbar $X(s) = K(s) - e$, da der Spieler ja $K(s)$ Euro gewinnt, aber einen Einsatz von e Euro zahlen musste. Anhand des Erwartungswertes sieht man, dass das Spiel für $e = 1/2$ fair ist:

$$E(X) \;=\; E(K-e) \;=\; E(K) - e \;=\; \sum_{k=0}^{3} kp(K=k) - e \;=\; \frac{1}{2} - e$$

Beispiel 7.14:
Eine Spielbank bietet folgendes Spiel an: Der Spieler wirft einen Würfel (d6), der Croupier wirft 2 Würfel und bildet deren Summe (2d6). Hat der Croupier ein (echt) höheres Ergebnis als der Spieler, muss der Spieler $v = 2$ Euro an die Bank zahlen. Hat der Spieler Glück und erreicht mindestens das Ergebnis des Croupiers, so erhält er $g = 10$ Euro von der Bank. Wir fragen uns:

(1) Wie hoch ist der vom Spieler erwartete Gewinn pro Spiel?

(2) Wie hoch müsste die Gewinnsumme g sein, damit das Spiel fair ist?

Für den Erwartungswert des Gewinns des Spielers benötigt man zunächst die Gewinnwahrscheinlichkeit $P = p(2d6 \leq d6)$. Diese lässt sich mit dem Satz der totalen Wahrscheinlichkeit berechnen:

$$p(2d6 \leq d6) \;=\; \sum_{s=2}^{12} p(s \leq d6 \mid 2d6 = s)\cdot p(2d6 = s)$$

Da die Würfel voneinander unabhängig sind, ist die bedingte Wahrscheinlichkeit $p(s \leq d6 \mid 2d6 = s) = p(s \leq d6) = (7-s)/6$ für $s \leq 6$ bzw. $= 0$ für $s > 6$. Die Wahrscheinlichkeit mit 2 Würfeln s zu würfeln beträgt (für $s \in 1 \ldots s$, siehe dazu die Tabelle in dem einführende Beispiel zu diesem Kapitel) $p(2d6 = s) = (s-1)/36$. Insgesamt also

$$P \;=\; p(2d6 \leq d6) \;=\; \sum_{s=2}^{6} \frac{7-s}{6}\cdot\frac{s-1}{36} \;=\; \frac{4}{27}$$

Die Wahrscheinlichkeit, dass der Spieler verliert, beträgt also 23/27. Und damit kann man den Erwartungswert des Gewinns des Spielers berechnen

$$E(\text{Gewinn}) \;=\; g\cdot P - v\cdot(1-P) \;=\; -\frac{2}{9}$$

Entsprechend erwartet die Bank 2/9 also etwa 22 Cent Gewinn pro Spiel. Das Spiel wäre *fair*, wenn $E(\text{Gewinn}) = 0$ ist. Löst man die entsprechende Gleichung $0 = g \cdot P - v \cdot (1 - P)$ für den Erwartungswert also nach g auf:

$$g \;=\; v \cdot \frac{1-P}{P} \;=\; 13.5$$

Bemerkung 7.15: Maximum Likelihood Methode:
Wir betrachten nun folgendes Problem: Oftmals kennt man nur die Art der Wahrscheinlichkeitsverteilung (z.B. eine Binomialverteilung bei Münzwürfen), nicht aber die konkreten Wahrscheinlichkeiten (etwa bei einer nicht idealen Münze). Man möchte daher die zugrunde liegenden Wahrscheinlichkeiten anhand von Messwerten $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$ schätzen.

Sei S also eine endliche (!) Menge und zu jedem Parameter $\lambda \in [0,1]$ sei $p_\lambda : S \to \mathbb{R}$ eine Wahrscheinlichkeitsverteilung, d.h. (S, p_λ) sei ein Laplace-Raum. Sei ferner $X : S \to \mathbb{R}$ eine Zufallsvariable. Dann ist

$$p_k(\lambda) \;:=\; p_\lambda(X = x_k) \;=\; \sum_{X(s)=x_k} p_\lambda(s)$$

die Wahrscheinlichkeit dafür, dass X den Wert x_k annimmt. Wir wollen nun λ so einstellen, dass die Wahrscheinlichkeiten $p_k(\lambda)$ möglichst groß werden. D.h. die *Likelihood-Funktion* $L(\lambda)$ soll möglichst groß werden:

$$L(\lambda) \;:=\; \prod_{k=1}^{n} p_k(\lambda)$$

Der sog. *Maximum Likelihood Estimate* $\textsc{mle}(X, x)$ von λ ist derjenige Parameterwert, für den $L(\lambda)$ maximal wird, d.h. für den gilt:

$$L(\textsc{mle}(X,x)) \;=\; \sup\{\, L(\lambda) \mid \lambda \in [0,1] \,\}$$

Man beachte, dass es aufgrund der strengen Monotonie des Logarithmus $\ln : \mathbb{R}_{>0} \to \mathbb{R}$ genügt, die logarithmierte Likelihood-Funktion zu maximieren:

$$\ell(\lambda) \;:=\; \ln(L(\lambda)) \;=\; \sum_{k=1}^{n} \ln(p_k(\lambda))$$

Das Maximum muss an einem kritischen Punkt von ℓ liegen, der Ansatz $\ell'(\lambda) = 0$ führt aber auf eine Gleichung, die oftmals lösbar ist:

$$\ell'(\lambda) \;=\; \sum_{k=1}^{n} \frac{p_k'(\lambda)}{p_k(\lambda)} \;=\; 0 \;\longrightarrow\; \lambda \;=\; \textsc{mle}(X,x)$$

Eine natürliche Verallgemeinerung dieser Situation besteht darin, dass eine endliche Anzahl von Parametern $\lambda = (\lambda_1, \ldots, \lambda_r)$ geschätzt werden muss.

Dies ist zum Beispiel der Fall, wenn Kugeln von $1 \leq r \in \mathbb{N}$ verschiedenen Sorten aus einer Urne gezogen werden. Will man die Anteile λ_i der Kugeln der Sorte $i \in 1 \ldots r$ an der Gesamtzahl der Kugeln in der Urne schätzen, so kommen nicht mehr alle λ in Frage, sondern es besteht die *Nebenbedingung*

$$\lambda \;\in\; \Delta_r \;:=\; \left\{ (\lambda_1, \ldots, \lambda_r) \in \mathbb{R}^r \;\middle|\; \begin{array}{l} \forall\, i \in 1 \ldots r : \lambda_i \geq 0 \\ \text{und } \lambda_1 + \cdots + \lambda_r = 1 \end{array} \right\}$$

Es sei also wieder S eine endliche Menge und zu jedem $\lambda = (\lambda_1, \ldots, \lambda_r) \in \Delta_r$ sei ein Laplace-Raum (S, p_λ) gegeben. Ist wieder $X : S \to \mathbb{R}$ eine Zufallsvariable und $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$ eine spezielle Messung von X, dann definieren wir wieder die logarithmierte *Likelihood-Funktion*

$$p_k(\lambda) \;:=\; p_\lambda(X = x_k) \;=\; \sum_{X(s)=x_k} p_\lambda(s)$$

$$\ell(\lambda) \;:=\; \sum_{k=1}^{n} \ln(p_k(\lambda))$$

Wir wollen $\ell(\lambda)$ wieder maximieren - dürfen dabei aber die Nebenbedingung $n(\lambda) := \lambda_1 + \ldots + \lambda_r - 1 = 0$ nicht verletzen. Nach dem Satz über Maxima unter Nebenbedingungen führen wir dazu einen *Lagrange-Multiplikator* $\mu \in \mathbb{R}$ und suchen einen kritischer Punkt der *Lagrange-Funktion* Λ

$$\Lambda(\lambda, \mu) \;:=\; \ell(\lambda) - \mu \cdot n(\lambda)$$

Die partielle Ableitung nach μ ist einfach die Nebenbedingung $n(\lambda) = 0$ selbst. Und die partiellen Ableitungen $\partial_i := \partial / \partial \lambda_i$ ergeben wegen $\partial_i n(\lambda) = 1$

$$0 \;=\; \partial_i \Lambda(\lambda, \mu) \;=\; \sum_{k=1}^{n} \frac{\partial_i p_k(\lambda)}{p_k(\lambda)} - \mu$$

Insgesamt ist also ein nicht-lineares Gleichungssystem [bestehend aus $r+1$ Gleichungen für die $r+1$ Unbekannten (λ, μ)] zu lösen, um die Wahrscheinlichkeit, dass der Laplace-Raum (S, p_λ) zu den Parametern $\lambda \in \Delta_r$ unter der Zufallsvariablen $X : S \to \mathbb{R}$ die Daten $x \in \mathbb{R}^n$ ergibt, zu maximieren:

$$\left.\begin{array}{rcl} \sum_{k=1}^{n} \frac{\partial_i p_k(\lambda)}{p_k(\lambda)} & = & \mu \\ \lambda_1 + \ldots + \lambda_r & = & 1 \end{array}\right\} \;\longrightarrow\; \lambda \;=\; \textsc{mle}(X, x)$$

Bislang haben wir die Nebenbedingungen $\lambda_i \geq 0$ aber noch nicht betrachtet. Will man diese mit einbeziehen, muss man weitere Multiplikatoren $\gamma = (\gamma_1, \ldots, \gamma_r) \in \mathbb{R}^r$ mit $\gamma_i \geq 0$ einführen. Die KKT-Bedingungen (siehe unten) zum Auffinden der optimalen Parameter λ bestehen dann in der Lösung eines nicht-linearen Gleichungssystems [bestehend aus $2r+1$ Gleichungen für die $2r+1$ Unbekannten (λ, γ, μ)]

$$\left.\begin{array}{rcl} \sum_{k=1}^{n} \frac{\partial_i p_k(\lambda)}{p_k(\lambda)} & = & \mu - \gamma_i \\ \gamma_i \lambda_i & = & 0 \\ \lambda_1 + \ldots + \lambda_r & = & 1 \end{array}\right\} \;\longrightarrow\; \lambda \;=\; \textsc{mle}(X, x)$$

Bemerkung 7.16: Maxima unter Nebenbedingungen:
Ohne Beweis möchten wir hier einen recht starken Satz für Maxima unter Nebenbedingungen angeben, von *Karush*, *Kuhn* und *Tucker*: Sei $D \subseteq \mathbb{R}^n$ und $f : D \to \mathbb{R}$. Weiterhin seien zu $i \in 1 \dots \ell$ und $j \in 1 \dots m$ die Funktionen $g_i : D \to \mathbb{R}$ und $n_j : D \to \mathbb{R}$ gegeben. Dabei seien f und alle g_i und alle n_j stetig differenzierbar auf einer offenen Umgebung von D. Bezeichne nun

$$N \quad := \quad \left\{ x \in D \;\middle|\; \begin{array}{l} \forall\, i \in 1 \dots \ell : g_i(x) \leq 0 \\ \forall\, j \in 1 \dots m : n_j(x) = 0 \end{array} \right\}$$

Dann suchen wir ein Minimum von $f(x)$ unter der Nebenbedingung $x \in N$, d.h. wir nehmen an, wir hätten einen Punkt $x_* \in N$ gegeben, so dass gilt:

$$f(x_*) \quad = \quad \min\{\, f(x) \mid x \in N \,\}$$

Schließlich bezeichnen wir zu $\lambda = (\lambda_1, \dots, \lambda_\ell) \in \mathbb{R}^\ell$ und $\mu = (\mu_1, \dots, \mu_m) \in \mathbb{R}^m$ die *Lagrange-Funktion*

$$L(x, \lambda, \mu) \quad := \quad f(x) + \sum_{i=1}^{\ell} \lambda_i g_i(x) + \sum_{j=1}^{m} \mu_j n_j(x)$$

Erfüllen die Nebenbedingungen nun die folgenden *Regularitätsbedingungen* in x_*: die folgende Teilmenge ist linear unabhängig in $\mathbb{R}^n$

$$\{\, \nabla n_j(x_*) \mid j \in 1 \dots m \,\} \cup \{\, \nabla g_i(x_*) \mid i \in 1 \dots \ell \;\text{ mit }\; g_i(x_*) = 0 \,\}$$

dann gibt es (nicht notwendiger Weise eindeutige) $\lambda_i \geq 0$ und $\mu_j \in \mathbb{R}$ (wobei $i \in 1 \dots \ell$ und $j \in 1 \dots m$), so dass die folgenden beiden Gleichungen gelten:

$$\begin{array}{rcl} \nabla_x L(x, \lambda, \mu) & = & 0 \\ \lambda_i g_i(x) & = & 0 \end{array}$$

Insgesamt ist also folgendes Gleichungssystem, bestehend aus $n + \ell + m$ Gleichungen und ebenso vielen Unbekannten (x, λ, μ) zu lösen, wobei wir zusätzlich wissen, dass $\lambda_i \geq 0$ und $x \in N$ sein kann:

$$\left.\begin{array}{rcl} \nabla_x L(x, \lambda, \mu) & = & 0 \\ \lambda_i g_i(x) & = & 0 \\ n_j(x) & = & 0 \end{array}\right\}$$

Kapitel 8

Spezielle Verteilungen

In diesem Abschnitt wollen wir einige klassische Verteilungen betrachten. In Kapitel 13 wird dies fortgesetzt und auch kontinuierliche Verteilungen betrachtet. Zunächst beschränken wir uns aber auf diskrete Verteilungen, die als Laplace-Raum formuliert werden können. Auf die Dauer müssen wir die Laplace-Räume aber hinter uns lassen um beispielsweise die Normalverteilung betrachten zu können. Doch zunächst eine kurze Übersicht über die in diesem Text dargestellten Verteilungen und deren Anwendung:

Verteilung	Symbol	modelliert
binomial	$\mathcal{B}(n,\lambda)$	n Würfe einer (2-seitigen) Münze
multinomial	$\mathcal{M}(n,\lambda)$	n Würfe eines r-seitigen Würfels
		$= n$ Ziehungen mit Zurücklegen
hypergeometrisch	$\mathcal{H}(n,m)$	n Ziehungen ohne Zurücklegen
geometrisch	$\mathcal{G}(\lambda)$	Anzahl Versuche bis zum Erfolg
Poisson	$\mathcal{P}(\lambda)$	Anzahl der Vorkommen eines seltenen
		Ereignisses über lange Zeiträume
normal	$\mathcal{N}(\mu,\sigma)$	entsteht bei der Überlagerung
		vieler unabhängiger Einflüsse
chi-Quadrat	$\mathcal{C}(n)$	testet die Güte eines Modells
		in Bezug auf vorliegende Daten

Binomialverteilung:
Wir betrachten ein Experiment mit zwei möglichen Ausgängen, z.B. das Werfen einer Münze mit den Ergebnissen $1 = \mathit{Kopf}$ oder $0 = \mathit{Zahl}$. Werfen wir $1 \leq n \in \mathbb{N}$ Mal und notieren die Ergebnisse, dann ist die Folge der Würfe also ein n-Tupel $s = (s_1, \ldots, s_n) \in S$ in $S = \{\,0,1\,\}^n$. Und für eine solche Wurffolge s sei $\omega(s) \in \mathbb{N}$ die Anzahl der Würfe, die 1 ergeben haben:

$$\omega(s) \;:=\; \#\{\,k \in 1\ldots n \mid s_k = 1\,\}$$

Ist $\lambda \in [0,1]$ die Wahrscheinlichkeit, bei einem Wurf eine 1 zu erhalten, dann ist die Wahrscheinlichkeit 0 zu erhalten also gerade $1-\lambda$. Und da die Ergebnisse der n verschiedenen Würfe voneinander unabhängig sind, ist die Wahrscheinlichkeit genau die Wurffolge s zu erhalten also $\lambda^k(1-\lambda)^{n-k}$ wobei $k = \omega(s)$ ist.

Es gibt aber $\binom{n}{k}$ verschiedene Wurffolgen $s \in S$, die alle genau $k \in 0 \dots n$ Mal 1 gezeigt haben. Insgesamt ist die Wahrscheinlichkeit k Mal die 1 zu werfen also gegeben durch

$$p(\omega = k) \quad = \quad \binom{n}{k} \lambda^k (1-\lambda)^{n-k}$$

Diese Überlegung führt auf die folgende Definition der Binomialverteilung. Eine Visualisierung dieses Experiments liefert zum Beispiel das Galton-Brett [siehe etwa http://de.wikipedia.org/wiki/Galtonbrett]. Dabei ist dann n die Zahl der Reihen mit Nägeln, die die Kugeln herunter fallen und λ die Wahrscheinlichkeit in einem beliebigen Schritt nach rechts zu fallen. Für genügend viele Kugeln ergibt sich dabei dann die Binomialverteilung.

Definition 8.1: Binomialverteilung
Wir definieren die **Binomialverteilung** der *Tiefe* $1 \leq n \in \mathbb{N}$ zum Parameter (zur Wahrscheinlichkeit) $\lambda \in [0,1]$ als den Laplace-Raum

$$\mathcal{B}(n, \lambda) \quad := \quad (S, p)$$

wobei $S = \{0, 1\}^n$ und zu jedem $s = (s_1, \dots, s_n) \in S$ bezeichnen wir die Zahl der 1en in s, mit $\omega(s) := \#\{k \in 1 \dots n \mid s_k = 1\}$. Dann definieren wir damit auch die Wahrscheinlichkeitsverteilung $p : S \to \mathbb{R}^+$ durch

$$p(s) \quad := \quad \lambda^{\omega(s)} (1-\lambda)^{n-\omega(s)}$$

Satz: Mit diesen Bezeichnungen ist für eine vorgelegte Zahl $k \in 0 \dots n$ die Wahrscheinlichkeit beim Wurf k 1en zu erhalten, gegeben durch

$$p(\omega = k) \quad = \quad b(n, \lambda, k) \quad := \quad \binom{n}{k} \lambda^k (1-\lambda)^{n-k}$$

Satz 8.2:

(i) Sei $(S, p) = \mathcal{B}(n, \lambda)$ die Binomialverteilung mit der Tiefe $n \in \mathbb{N}$ zum Parameter $\lambda \in [0,1]$, und wie in der obigen Definition sei wieder $\omega(s) := \#\{k \in 1 \dots n \mid s_k = 1\}$, dann gilt für Erwartungswert und Varianz

$$\begin{aligned} E(\omega) &= n\lambda \\ V(\omega) &= n\lambda(1-\lambda) \end{aligned}$$

Allgemeiner gilt für die Momente-erzeugende Funktion (ein Begriff, der in Kapitel 11 eingeführt wird) bei der Binomialverteilung

$$M_\omega(t) \quad = \quad (\lambda e^t - (1-\lambda))^n$$

(ii) Bezeichnen wir den Erwartungswert mit $A := E(\omega)$ und die Varianz mit $V := V(\omega)$ dann können wir für $\lambda \neq 0$ die Tiefe und den Parameter aus A und V zurück gewinnen, durch

$$\begin{aligned} \lambda &= \frac{A-V}{A} \\ n &= \frac{A^2}{A-V} \end{aligned}$$

(iii) Ist $x = (x_1, \ldots, x_m) \in (0 \ldots n)^m$ ein Datentupel (x_i ist die Position der i-ten Kugel nach dem Durchlaufen eines n-stufigen Galton-Brettes), dann ist die Maximum Likelihood Schätzung des Parameters λ das Tupel x unter ω zu beobachten, gegeben durch

$$\textsc{mle}(\omega, x) = \frac{1}{n} A(x)$$

(v) Der Parameter λ der Binomialverteilung kann auch direkt aus den Werten der Verteilung ermittelt werden: bezeichnen wir zu $k \in 0 \ldots n$ jeweils $n_k := b(n, \lambda, k)$ dann gilt

$$\lambda = \frac{k n_k}{(n-k+1) n_{k-1} + k n_k}$$

(vi) Zu jedem Parameterwert $\lambda \in [0, 1]$ und jedem $k \in 0 \ldots (n-1)$ erfüllen die Symbole $b(n, \lambda, k)$ die folgende Rekursionsformel

$$b(n, \lambda, k+1) = \frac{n-k}{k+1} \cdot \frac{\lambda}{1-\lambda} \cdot b(n, \lambda, k)$$

(vi) Die Gegenwahrscheinlichkeit der kumulierten Verteilung lässt sich durch die kumulierte Verteilung der Gegenwahrscheinlichkeit ausdrücken. Konkret gilt für beliebiges $\lambda \in [0, 1]$ und $k \in 0 \ldots n$ gilt die Formel

$$1 - B(n, \lambda, k-1) = B(n, 1-\lambda, n-k)$$

(vii) *Chernoff-Schranke*: [John Canny, berkley.edu, CS174] Ist wieder $(S, p) = \mathcal{B}(n, \lambda)$, $\mu := E(\omega) = n\lambda$ und ist $0 < \delta < 1$, dann gilt die Abschätzung

$$p(\omega < (1-\delta)\mu) < \left(\frac{e^{-\delta}}{(1-\delta)^{1-\delta}} \right)^{\mu} < \exp\left(-\frac{\delta^2}{2} \mu \right)$$

Und ist $0 < \delta$ beliebig, dann erhalten wir die umgekehrte Abschätzung

$$p(\omega > (1+\delta)\mu) < \left(\frac{e^{\delta}}{(1+\delta)^{1+\delta}} \right)^{\mu} < \exp\left(-\frac{\delta^2}{2+\delta} \mu \right)$$

(viii) Die Verteilungsfunktion der Binomialverteilung ist eine sogenannte *Beta-Verteilung*. Sind $1 \leq a,\ b \in \mathbb{N}$ und ist $x \in [0,1]$ dann ist diese definiert, durch

$$\beta(x,a,b) \ := \ \frac{(a+b-1)!}{(a-1)! \cdot (b-1)!} \int_0^x t^{a-1}(1-t)^{b-1} dt$$

Diese kann noch weiter verallgemeinert werden, worauf wir hier aber nicht eingehen wollen. In dieser Form wird sie aber von MS-Excel oder Libre-Office-Calc zur Verfügung gestellt. So berechnet man $\beta(x,a,b)$ mit dem Befehl $\mathsf{BETADIST}(x,a,b,0,1)$. Und sogar die Inverse Beta-Verteilung kann berechnet werden: Ist $y = \beta(x,a,b)$, dann liefert $x = \mathsf{BETAINV}(y,a,b,0,1)$ gerade die Variable x zurück. Mit dieser Funktion gilt dann für alle $0 \leq k < n$:

$$\sum_{i=0}^{k} b(n,\lambda,i) \ = \ \beta(1-\lambda, n-k, k+1)$$

Explizit ausgeschrieben bedeutet diese Identität also nichts anderes, als die folgende Möglichkeit zur Berechnung der ersten k Wahrscheinlichkeiten $b(n,\lambda,i)$

$$\sum_{i=0}^{k} \binom{n}{i} \lambda^i (1-\lambda)^{n-i} \ = \ \frac{n!}{(n-k-1)! \cdot k!} \int_0^{1-\lambda} t^{n-k-1}(1-t)^k dt$$

(ix) Seien $1 \leq m,\ n \in \mathbb{N}$ die Tiefen zweier Binomialverteilungen zu dem Parameter $\lambda \in [0,1]$. Dann ist die Binomialverteilung der Tiefe $m+n$ einfach das Produkt der Verteilungen - vgl. (11.7) - der Tiefe m und n

$$\mathcal{B}(m,\lambda) \otimes \mathcal{B}(n,\lambda) \ = \ \mathcal{B}(m+n,\lambda)$$

Beweis:
Wir wollen hier nur die erste Aussage (i) zeigen - die Formeln für den Erwartungswert und die Varianz der Binomialverteilung. Die Rechnungen in denen dies gezeigt wird sind elementar und länglich sollten aber der Vollständigkeit halber nicht fehlen. Wir betrachten nun eine Funktion in den Variablen x und y und bezeichnen die Ableitung nach x mit ∂_x

$$\begin{aligned}
f(x,y) \ &:= \ \sum_{k=0}^{n} k \binom{n}{k} x^k y^{n-k} \ = \ \sum_{k=1}^{n} k \binom{n}{k} x \cdot x^{k-1} y^{n-k} \\
&= \ x \cdot \sum_{k=1}^{n} \binom{n}{k} \partial_x (x^k) y^{n-k} \ = \ x \cdot \partial_x \sum_{k=1}^{n} \binom{n}{k} x^k y^{n-k} \\
&= \ x \cdot \partial_x \left(\sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k} - y^n \right) \ = \ x \cdot \partial_x \left((x+y)^n - y^n \right) \\
&= \ x \cdot n \cdot (x+y)^{n-1}
\end{aligned}$$

Nach Definition des Erwartungswertes ist $E(\omega)$ gerade $f(\lambda, 1-\lambda)$ und damit erhält man dann, die in (i) behauptete Formel:

$$E(\omega) \;=\; \sum_{k=0}^{n} k \cdot b(n,\lambda,k) \;=\; f(\lambda, 1-\lambda) \;=\; \lambda \cdot n \cdot 1^{n-1} \;=\; n\lambda$$

Für die Formel der Varianz verwenden wir den Verschiebungssatz $V(\omega) = E(\omega^2) - E(\omega)^2$ und zur Berechnung von $E(\omega^2)$ betrachten wir die Funktion:

$$\begin{aligned}
g(x,y) \;&:=\; \sum_{k=0}^{n} k^2 \binom{n}{k} x^k y^{n-k} \\
&=\; \sum_{k=1}^{n} k(k-1) \binom{n}{k} x^k y^{n-k} + \sum_{k=0}^{n} k \binom{n}{k} x^k y^{n-k} \\
&=\; \sum_{k=2}^{n} k(k-1) \binom{n}{k} x^k y^{n-k} + f(x,y) \\
&=\; x^2 \cdot \sum_{k=2}^{n} k(k-1) \binom{n}{k} x^{k-2} y^{n-k} + f(x,y) \\
&=\; x^2 \cdot \sum_{k=2}^{n} \binom{n}{k} \partial_x^2 (x^k) y^{n-k} + f(x,y) \\
&=\; x^2 \partial_x^2 \left(\sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k} - y^n - nxy^{n-1} \right) + f(x,y) \\
&=\; x^2 \partial_x^2 \left((x+y)^n - y^n - nxy^{n-1} \right) + nx(x+y)^{n-1} \\
&=\; x^2 \partial_x \left(n(x+y)^{n-1} - ny^{n-1} \right) + nx(x+y)^{n-1} \\
&=\; x^2 n(n-1)(x+y)^{n-2} + nx(x+y)^{n-1}
\end{aligned}$$

Wiederum ist g passend definiert worden, dass $E(\omega^2)$ gerade $g(\lambda, 1-\lambda)$ ist. Und $E(\omega)^2 = (n\lambda)^2$ kennen wir bereits nach dem oben Gezeigtem. Damit

$$\begin{aligned}
V(\omega) \;&=\; E(\omega^2) - E(\omega)^2 \;=\; g(\lambda, 1-\lambda) - (n\lambda)^2 \;=\; \lambda^2 n(n-1) + n\lambda - n^2\lambda^2 \\
&=\; n^2\lambda^2 - n\lambda^2 + n\lambda - n^2\lambda^2 \;=\; n\lambda - n\lambda^2 \;=\; n\lambda(1-\lambda)
\end{aligned}$$

Der Beweis von (vi) ist sehr einfach und soll hier ebenfalls nicht fehlen. In der Rechnung wird der Index $j = n - i$ eingeführt. Dann ist der Bereich $i \in k \ldots n$ gerade $j \in 0 \ldots k$ und daher gilt:

$$\begin{aligned}
1 - B(n,\lambda,k-1) \;&=\; \sum_{i=k}^{n} \binom{n}{i} \lambda^i (1-\lambda)^{n-i} \\
&=\; \sum_{i=k}^{n} \binom{n}{n-i} (1-\lambda)^{n-i} \lambda^{n-(n-i)} \\
&=\; \sum_{j=0}^{k} \binom{n}{j} (1-\lambda)^j \lambda^{n-j} \;=\; B(n, 1-\lambda, k)
\end{aligned}$$

□

Multinomialverteilung:
Wir verallgemeinern nun die Binomialverteilung auf mehr als 2 mögliche Ergebnisse des Zufallsexperiments: d.h. wir werfen einen r-seitigen Würfel n Mal und notieren uns die Ergebnisse $s = (s_1, \ldots, s_n) \in (1 \ldots r)^n$ der Würfelwürfe. Anstelle eines Würfels kann man auch ein Glücksrad betrachten: Ist λ_i der Anteil 'Winkel des i-ten Feldes / 2π' dann ist $\lambda = (\lambda_1, \ldots, \lambda_r)$ eine Wahrscheinlichkeitsverteilung auf $1 \ldots r$, d.h. es gilt $\lambda_i \geq 0$ und $\lambda_1 + \ldots + \lambda_r = 1$.

Dieselbe Situation erhält man, wenn man eine Urne mit r Sorten von Kugeln füllt. Sind jeweils $m_i \in \mathbb{N}$ Kugeln der i-ten Sorte gegeben, dann sind insgesamt $M = m_1 + \cdots + m_r$ Kugeln in der Urne. Die Wahrscheinlichkeit eine Kugel der Sorte i zu ziehen ist in diesem Fall also $\lambda_i = m_i/M$ und wieder ist $\lambda = (\lambda_1, \ldots, \lambda_r)$ eine Wahrscheinlichkeitsverteilung auf $1 \ldots r$.

Ist nun $s = (s_1, \ldots, s_n) \in (1 \ldots r)^n$ ein n-Tupel von Ergebnissen unseres Zufallsexperiments dann bezeichnen wir mit $\omega_i := \#\{\, k \in 1 \ldots n \mid s_k = i \,\}$ die Zahl der Vorkommen der Sorte i in den Ergebnissen. Und damit ist dann die Wahrscheinlichkeit *genau diese Ziehung* s zu erhalten gegeben, durch:

$$p(s) \quad = \quad \prod_{i=1}^{r} \lambda_i^{\omega_i(s)}$$

Bei der Zählung, welche Sorte wie oft vorkommt, spielt die Reihenfolge der Ergebnisse aber keine Rolle. Wir können die s_k beliebig umsortieren, wobei die $\omega_i(s)$ Ergebnisse der Sorte i dabei keinen Unterschied machen, so dass es (vgl. Kugelverteilung I) $\binom{n}{\alpha}$ Möglichkeiten gibt, die Anzahlen $\alpha_i := \omega_i(s) \in 0 \ldots n$ (wobei $i \in 1 \ldots r$) zu erhalten. Insgesamt beträgt die Wahrscheinlichkeit die Anzahlen $\alpha = (\alpha_1, \ldots, \alpha_r)$ zu erhalten also

$$p(\omega = \alpha) \quad = \quad \binom{n}{\alpha} \prod_{i=1}^{r} \lambda_i^{\omega_i(s)}$$

Definition 8.3: Multinomialverteilung:
Wir definieren die **Multinomialverteilung** mit r *Sorten*, der *Tiefe* n und den *Häufigkeiten* (auch *Parameter* genannt) $\lambda = (\lambda_1, \ldots, \lambda_r)$ (wobei $1 \leq r$, $n \in \mathbb{N}$ und für $i \in 1 \ldots r$ ist $\lambda_i \in [0, 1]$ so dass $\lambda_1 + \cdots + \lambda_r = 1$) als den folgenden Laplace-Raum

$$\mathcal{M}(n, \lambda) \quad := \quad (S, p)$$

wobei $S := (1 \ldots r)^n$ und die Wahrscheinlichkeitsverteilung p wie folgt definiert wird: zunächst bezeichnen wir zu $s = (s_1, \ldots, s_n) \in S$ die Anzahlen

$$\omega_i \ : \ S \to \mathbb{N} \ : \ (s_1, \ldots, s_n) \mapsto \#\{\, k \in 1 \ldots n \mid s_k = i \,\}$$

und schreiben $\omega(s) := (\omega_1(s), \ldots, \omega_r(s)) \in \mathbb{N}^r$, dann definieren wir p durch:

$$p(s) \quad := \quad \prod_{i=1}^{r} \lambda_i^{\omega_i(s)} \quad \in \quad \mathbb{R}^+$$

Satz: Bezeichnen wir die Menge alle Multi-indices mit Betrag n mit $A(r,n) := \{\, \alpha = (\alpha_1, \ldots, \alpha_r) \in \mathbb{N}^r \mid |\alpha| = \alpha_1 + \ldots + \alpha_r = n \,\}$, dann ist für $\alpha \in A(r,n)$ die Wahrscheinlichkeit ein $s \in S$ mit $\omega(s) = \alpha$ zu erhalten gegeben, durch:

$$p(\omega = \alpha) \;=\; m(n, \lambda, \alpha) \;:=\; \binom{n}{\alpha} \prod_{i=1}^{r} \lambda_i^{\alpha_i}$$

Satz 8.4:
Sei $(S,p) = \mathcal{M}(n,\lambda)$ die Multinomialverteilung zur Tiefe n mit den Häufigkeiten $\lambda_1, \ldots, \lambda_r$ der r Sorten. Und zu $i \in 1 \ldots r$ bezeichne wie üblich $\omega_i : S \to \mathbb{N} : (s_1, \ldots, s_n) \mapsto \#\{\, k \in 1 \ldots n \mid s_k = i \,\}$ die Zufallsvariable des Vorkommens der i-ten Sorte. Dann gilt:

(i) Ist $i \in 1 \ldots r$ und $k \in 0 \ldots n$, dann ist die Wahrscheinlichkeit für $\omega_i = k$ gerade die Wahrscheinlichkeit der Binomialverteilung mit n Stufen zu λ_i des Ereignisses $\omega = k$, d.h. es gilt

$$p(\omega_i = k) \;=\; b(n, \lambda_i, k) \;=\; \binom{n}{k} \lambda_i^k (1 - \lambda_i)^{n-k}$$

(ii) Für beliebige Sorten i, $j \in 1 \ldots r$ mit $i \neq j$ können wir den Erwartungswert, die Varianz und Kovarianz der ω_i und ω_j berechnen, durch

$$\begin{aligned} E(\omega_i) &= n\lambda_i \\ V(\omega_i) &= n\lambda_i(1 - \lambda_i) \\ \sigma(\omega_i \mid \omega_j) &= -n\lambda_i\lambda_j \end{aligned}$$

(iii) Ist $X = (x_1, \ldots, x_r)$ eine $(m \times r)$-Matrix, wobei $x_j = (x_{1,j}, \ldots, x_{m,j}) \in \mathbb{N}^m$ und $x_{i,1} + \ldots + x_{i,r} = n$ für alle $i \in 1 \ldots m$ ($x_{i,j}$ ist die Anzahl der Vorkommen der j-ten Sorte in der i-ten Ziehung), dann ist die Maximum Liklihood Schätzung der Parameter λ_j die Matrix X unter $\omega = (\omega_1, \ldots, \omega_r)$ zu beobachten, gegeben durch

$$\textsc{mle}(\omega, X) \;=\; \frac{1}{n}(A(x_1), \ldots, A(x_r))$$

(iv) Seien $1 \leq m$, $n \in \mathbb{N}$ die Tiefen zweier Multinomialverteilungen mit r Sorten, zu den selben Parametern $\lambda = (\lambda_1, \ldots, \lambda_r)$. Dann ist die Multinomialverteilung der Tiefe $m+n$ einfach das Produkt der Verteilungen - vgl. (11.7) - der Tiefe m und n

$$\mathcal{M}(m, \lambda) \otimes \mathcal{M}(n, \lambda) \;=\; \mathcal{M}(m + n, \lambda)$$

Hypergeometrische Verteilung:
Die Multinomialverteilung $\mathcal{M}(n, \lambda)$ beschreibt den Fall, dass aus einer Urne n Kugeln gezogen (und zurück gelegt) werden. Dabei gibt es r verschiedene Sorten von Kugeln, mit den Anteilen $\lambda_i = m_i/M$ an der Gesamtzahl M von Kugeln in der Urne. D.h. es sind m_i Kugeln der Sorte i in der Urne und

$$M = \sum_{i=1}^{r} m_i$$

Da wir die Kugeln zurück legen ändern sich die Anteile λ_i nicht und der n-te Zug ist unabhängig von allen voran gegangenen Zügen. Wenn wir die Kugeln jedoch *nicht zurücklegen*, ändern sich die Anteile und damit die Wahrscheinlichkeiten:

Nehmen wir an, wir hätten eine Kugel der Sorte i gezogen. Da sich die Zahl der verbleibenden Kugeln in der Urne geändert hat, fällt die Wahrscheinlichkeit eine weitere Kugel der Sorte i zu ziehen auf $(m_i - 1)/(M - 1)$ wohingegen die Wahrscheinlichkeit eine Kugel der Sorte $j \neq i$ zu ziehen auf $m_j/(M - 1)$ gestiegen ist. Insgesamt ist die Wahrscheinlichkeit der Zugfolge (i, i) bzw. (i, j) also gegeben, durch

$$p(i,i) = \frac{m_i}{M} \cdot \frac{m_i - 1}{M - 1} \quad \text{bzw.} \quad p(i,j) = \frac{m_i}{M} \cdot \frac{m_j}{M - 1}$$

Offensichtlich lässt sich diese Überlegung auf n aufeinanderfolgende Züge verallgemeinern: Nehmen wir an, wir hätten α_i Kugeln der Sorte i gezogen (für alle $i \in 1 \ldots r$). Dabei muss natürlich $\alpha_i \leq m_i$ sein und es ist

$$n = |\alpha| = \sum_{i=1}^{r} \alpha_i$$

Da sich die Gesamtzahl der Kugeln in der Urne in jedem Zug um eins verringert, wird nach n Zügen der Nenner in dieser Kette von Anteilen angewachsen sein zu

$$\frac{1}{M} \cdot \frac{1}{M - 1} \cdot \ldots \cdot \frac{1}{M - n + 1} = \frac{(M - n)!}{M!}$$

Und die α_i Ziehungen von Kugeln der Sorte i würden zu den Zählern in der Kette der Anteile den Beitrag $m_i(m_i - 1) \ldots (m_i - \alpha_i + 1) = m_i!/(m_i - \alpha_i)!$ leisten. Da die Wahrscheinlichkeiten nur von den gegenwärtigen Anteilen der Kugeln (der jeweiligen Sorte) abhängen und diese Verhältnisse untereinander kommutieren, ist die Wahrscheinlichkeit α_1 Kugeln der Sorte 1 *und* α_2 Kugeln der Sorte 2 *und* und so weiter *und* α_r Kugeln der Sorte r zu ziehen, gegeben durch

$$p(s) = \prod_{i=1}^{r} \frac{m_i!}{(m_i - \alpha_i)!} \cdot \frac{(M - n)!}{M!} = \frac{\prod_{i=1}^{r} \alpha_i! \binom{m_i}{\alpha_i}}{n! \binom{M}{n}} = \frac{\prod_{i=1}^{r} \alpha_i!}{n! \binom{M}{n}} \prod_{i=1}^{r} \binom{m_i}{\alpha_i}$$

wobei $s \in (1 \dots r)^n$ die Folge der Sorten der gezogenen Kugeln ist. Man beachte, dass die Wahrscheinlichkeit s zu ziehen in Wirklichkeit nur von den Anzahlen α_i von gezogenen Kugeln der Sorte i abhängt, nicht von der tatsächlichen Reihenfolge der Ziehung. Es gibt genau $n!/(\alpha_1! \dots \alpha_r!)$ Zugfolgen $s \in (1 \dots r)^n$ die die gemeinsamen Anzahlen $\alpha = (\alpha_1, \dots, \alpha_r)$ an Kugeln der verschiedenen Sorten aufweisen. Die Wahrscheinlichkeit insgesamt also α_1 Kugeln der Sorte 1 und α_2 Kugeln der Sorte 2 und so weiter bis α_r Kugeln der Sorte r zu ziehen, ist also gegeben, durch

$$p(\omega = \alpha) \quad = \quad \frac{1}{\binom{M}{n}} \prod_{i=1}^{r} \binom{m_i}{\alpha_i}$$

Beispiel 8.5:
Wir betrachten ein Beispiel mit $r = 2$ Sorten: Wie hoch ist die Wahrscheinlichkeit beim Lotto 4 Richtige (wir ignorieren die Zusatzzahl) zu haben? Dazu drehen wir die zeitliche Abfolge um: *Nach* der Ziehung gibt es 6 richtige und $49 - 6 = 43$ falsche (nicht-gezogene) Kugeln, es ist also $m = (6, 43)$. Beim Ausfüllen des Lottoscheins haben wir ebenfalls $n = 6$ Kreuzchen auf den 49 Feldern gemacht.

Die Frage ist also, wie viele Kreuzchen auf den richtigen 6 Feldern gelandet sind. Bei 4 Richtigen wären offenbar 2 in den falschen Feldern gelandet, so dass $\alpha = (4, 2)$. Die Wahrscheinlichkeit für diese Kreuzchenwahl ist also gegeben, durch

$$p(\omega = \alpha) \quad = \quad \frac{\binom{6}{4}\binom{43}{2}}{\binom{49}{6}} \quad \approx \quad 0.097\%$$

Definition 8.6: Hypergeometrische Verteilung
Es seien $1 \leq r$, $n \in \mathbb{N}$ und zu jedem $i \in 1 \dots r$ sei $m_i \in \mathbb{N}$ gegeben. Wir bezeichnen $m = (m_1, \dots, m_r) \in \mathbb{N}^r$ und definieren die **hypergeometrische Verteilung** mit r *Sorten*, der *Tiefe* n und den *Vorkommen* m, als den folgenden Laplace-Raum

$$\mathcal{H}(n, m) \quad := \quad (S, p)$$

Zunächst konstruieren wir S: Sei $T := (1 \dots r)^n$ und wie zuvor, die bei der Multinomialverteilung definieren Anzahlen ω_i der Vorkommen von i in s

$$\omega_i \ : \ T \to \mathbb{N} \ : \ (s_1, \dots, s_n) \mapsto \#\{\, k \in 1 \dots n \mid s_k = i \,\}$$

Wir bezeichnen ferner $M := m_1 + \dots + m_r \in \mathbb{N}$ und $\omega = (\omega_1, \dots, \omega_r) : T \to \mathbb{N}^r$. Dann können wir den Raum S beschreiben, als

$$S \quad := \quad \{\, s \in T \mid \forall\, i \in 1 \dots r : \omega_i(s) \leq m_i \,\}$$

Es bleibt die Wahrscheinlichkeitsverteilung $p : S \to \mathbb{R}^+$ zu definieren, durch:

$$p(s) \quad := \quad \frac{\prod_{i=1}^{r} \omega_i(s)!}{n! \binom{M}{n}} \prod_{i=1}^{r} \binom{m_i}{\omega_i(s)}$$

Satz: Bezeichnen wir die Menge aller Multi-indices $\alpha = (\alpha_1, \ldots, \alpha_r) \in \mathbb{N}^r$ mit Betrag n und $\alpha_i \leq m_i$ als $A_m(r,n) := \{\, \alpha \in A(r,n) \mid \forall\, i \in 1 \ldots r : \alpha_i \leq m_i \,\}$ dann ist für $\alpha \in A_m(r,n)$ die Wahrscheinlichkeit ein $s \in S$ mit $\omega(s) = \alpha$ zu erhalten gegeben, durch:

$$p(\omega = \alpha) \;:=\; h(n,m,\alpha) \;:=\; \frac{1}{\binom{M}{n}} \prod_{i=1}^{r} \binom{m_i}{\alpha_i}$$

Satz 8.7:
Sei $(S,p) = \mathcal{H}(n,m)$ die hypergeometrische Verteilung mit r Sorten und Tiefe n zum Vorkommen $m = (m_1, \ldots, m_r) \in \mathbb{N}^r$. Wie oben definieren wir $M := m_1 + \ldots + m_r$ und betrachten $i \neq j \in 1 \ldots r$, dann erhalten wir die folgenden Eigenschaften:

$$\begin{aligned} E(\omega_i) &= \frac{nm_i}{M} \\ V(\omega_i) &= \frac{nm_i}{M}\left(1 - \frac{m_i}{N}\right)\frac{M-n}{M-1} \\ \sigma(\omega_i \mid \omega_j) &= -\frac{nm_im_j}{M^2}\frac{M-n}{M-1} \end{aligned}$$

Bemerkung 8.8:
Wenn so wenige n Ziehungen erfolgen, dass sich die Anteile m_i/M nur wenig ändern, dann liefert die hypergeometrische Verteilung $\mathcal{H}(n,m)$ näherungsweise dieselben Werte, wie die Multinomialverteilung $\mathcal{M}(n,\lambda)$, d.h. für alle $\alpha \in A_m(r,n)$ gilt (wobei $M := m_1 + \ldots + m_r$ und $\lambda_i = m_i/M$)

$$h(n,m,\alpha) \;\approx\; m(n,\lambda,\alpha)$$

Genauer gesagt, gibt es aufgrund der Stirling'schen Formeln, Näherungsfehler ε und $\delta_i \in \mathbb{R}$ mit $|\delta_i| < 1/11(m_i - \alpha_i)$ beziehungsweise $|\varepsilon| < 1/11(M-n)$ so dass sich das Verhältnis dieser beiden Verteilungen schreiben lässt, als

$$\frac{h(n,m,\alpha)}{m(n,\lambda,\alpha)} \;=\; \frac{\prod_{i=1}^{r}(1+\delta_i)}{1+\varepsilon}\left(\frac{M-n}{M}\right)^{M-n+0.5} \prod_{i=1}^{r}\left(\frac{m_i}{m_i-\alpha_i}\right)^{m_i-\alpha_i+0.5}$$

Aufgrund der Komplexität dieser Formel ist es schwer eine vernünftige Abschätzung zu geben, wann ein Übergang zur Multionomialverteilung gerechtfertigt ist. In der Literatur wird gelegentlich das Kriterium $\alpha_i/m_i \leq 1/20$ (für alle $i \in 1 \ldots r$) gegeben. In diesem Fall sind dann auch $n/M \leq 1/20$ und $m_i/(m_i - \alpha_i) \leq 20/19$. Aber selbst in diesem Fall ist die Abschätzung noch recht unhandlich

$$\frac{h(n,m,\alpha)}{m(n,\lambda,\alpha)} \;\leq\; \frac{1+\sum_{i=1}^{r}\delta_i + \mathcal{O}(\delta_i^2)}{1+\varepsilon}\left(\frac{20}{19}\right)^{M-n+\frac{r}{2}}\left(\frac{M-n}{M}\right)^{M-n+0.5}$$

Geometrische Verteilung:
Wir kehren zur allgemein beliebten Binomialverteilung zurück - d.h. wir betrachten ein Bernoulli-Experiment, wie etwa die Frage ob ein geworfener Würfel eine 6 zeigt. Die Wahrscheinlichkeit, dass das gewünschte Ereignis eintritt sei $\lambda \in [0,1]$. Diesmal fragen wir uns aber, wie oft wir den Versuch wohl durchführen müssen, bis das Ereignis eintritt? Offensichtlich ist $p(0) = 0$ - wirft man den Würfel nicht, wird auch keine 6 angezeigt. Und kommt erst beim n-ten Mal eine 6, dann müssen zuvor $n-1$ Würfe jeweils eine Zahl von 1 bis 5 ergeben haben. Für $n \geq 1$ ist die Wahrscheinlichkeit erst nach n Würfen den ersten Erfolg zu haben, also gegeben durch

$$p(n) \quad = \quad (1-\lambda)^{n-1} \cdot \lambda$$

Es handelt sich hier um die einfachste Verteilung überhaupt - summiert man über alle $p(n)$, so entsteht eine geometrische Reihe (daher auch der Name, die Verteilung hat keine geometrische Bedeutung). Ist die Wartezeit nicht in diskreten Schritten, sondern kontinuierlich, dann muss man stattdessen die Poisson-Verteilung verwenden, die wir im Anschluss einführen werden.

Definition 8.9: Geometrische Verteilung
Sei $\lambda \in [0,1]$ ein fixierter Parameter (die Wahrscheinlichkeit des Eintretens des betrachteten Ereignis), dann definieren wir die **geometrische Verteilung** zu λ also den Laplace Raum

$$\mathcal{G}(\lambda) \quad := \quad (\mathbb{N}, p)$$

wobei wir $p(0) := 0$ setzen und zu $n \neq 0$ definieren wir die Wahrscheinlichkeit

$$p(n) \quad := \quad \lambda \cdot (1-\lambda)^{n-1}$$

Satz 8.10:
Sei $(\mathbb{N}, p) = \mathcal{G}(\lambda)$ die geometrische Verteilung zu $\lambda \in [0,1]$. Die Zufallsvariable der Anzahl der Versuche ist dann $\mathrm{id} : \mathbb{N} \to \mathbb{N} : n \to n$ und für diese gelten die folgenden Aussagen:

(i) Die Wahrscheinlichkeit, dass das betrachtete Ereignis nach *höchstens* $n \in \mathbb{N}$ Versuchen eintritt, ist gegeben, durch

$$p(\mathrm{id} \leq n) \quad = \quad 1 - (1-\lambda)^n$$

(ii) Ist $\lambda > 0$, dann hat id einen Erwartungswert und eine Varianz, von

$$\begin{aligned} E(\mathrm{id}) &= \frac{1}{\lambda} \\ V(\mathrm{id}) &= \frac{1-\lambda}{\lambda^2} \end{aligned}$$

Allgemeiner gilt für die momente-erzeugende Funktion (ein Begriff, der in Kapitel 11 eingeführt wird) bei der geometrischen Verteilung

$$M_{\mathrm{id}}(t) \quad = \quad \frac{\lambda t}{1-(1-\lambda)\cdot t}$$

(iii) Sei $x = (x_1, \ldots, x_n) \in \mathbb{N}^n$ mit $x_i \geq 1$ (für $i \in 1 \ldots n$) ein Datentupel (x_i ist die Anzahl der Versuche bis zum ersten Eintritt des gesuchten Ereignis), dann ist die Maximum Likelihood Schätzung des Parameters λ das Tupel x unter id zu beobachten, gegeben durch

$$\textsc{mle}(\mathrm{id}, x) \quad = \quad \frac{1}{A(x)}$$

Poisson-Verteilung:
Die Poisson-Verteilung ist eine Variante der Binomialverteilung: Sie modelliert die Situation, dass ein bestimmtes Ereignis eine konstante (geringe) Eintrittswahrscheinlichkeit hat und wir fragen uns wie häufig dieses Ereignis dann auf lange Sicht vorkommt. Ein Beispiel: Eine Folge von Bits wird mit konstanter Geschwindigkeit über eine Leitung versendet. Die Wahrscheinlichkeit, dass ein falsches Bit übertragen wird ist, ist proportional zur Zahl der übertragenen Bits und daher zur vergangenen Zeit. Wir fragen uns nun, wie hoch die Wahrscheinlichkeit ist, in einem langen Datenstrom ist, dass $s \in \mathbb{N}$ Bits falsch übertragen wurden? Wir treffen also folgende Annahmen:

- Die Wahrscheinlichkeit für den Eintritt des Ereignisses ist proportional zur betrachteten Zeitdauer Δt.
- Die Wahrscheinlichkeit für den Eintritt des Ereignisses ist unabhängig vom betrachteten Zeitpunkt t - nur von der Zeitspanne Δt.
- Das Ereignis ist *selten* - es kommt maximal einmal pro Zeitintervall Δt vor - da wir später $\Delta t \to 0$ gehen lassen, ist diese Annahme vernünftig.

Sei $p_0(t)$ die Wahrscheinlichkeit, dass das Ereignis im Zeitabschnitt $[0, t]$ nicht aufgetreten ist. Dann ist $p_0(t + \Delta t)$ die Wahrscheinlichkeit, dass das Ereignis im verlängerten Zeitintervall $[0, t+\Delta t]$ immer noch nicht aufgetreten ist. Dazu darf es also nicht im Zeitabschnitt $[0, t]$ *und* nicht im Zeitabschnitt $[t, t + \Delta t]$ aufgetreten sein. Da wir vorausgesetzt haben, dass die Eintrittswahrscheinlichkeit zeitunabhängig ist, ist letzteres also $p_0(\Delta t)$. Und da beide unabhängig voneinander mit *und* kombiniert werden, gilt also

$$p_0(t + \Delta t) \quad = \quad p_0(t) \cdot p_0(\Delta t)$$

Ist Δt klein genug, so tritt das Ereignis höchstens einmal im Zeitabschnitt $[0, \Delta t]$ ein. Ist $p_1(t)$ also die Wahrscheinlichkeit für das einmalige Eintreten des Ereignisses im Zeitabschnitt $[0, t]$, so gilt also $p_1(\Delta t) = 1 - p_0(\Delta t)$. Nun ist die Eintrittswahrscheinlichkeit aber proportional zur Zeitdauer, d.h. es ist $p_1(\Delta t) = \alpha \cdot \Delta t$ für eine Wahrscheinlichkeitsdichte $\alpha \geq 0$. Daraus folgt dann aber $p_0(\Delta t) = 1 - p_1(\Delta t) = 1 - \alpha \Delta t$ und damit

$$p_0(t + \Delta t) \;=\; p_0(t) \cdot (1 - \alpha \Delta t) \;=\; p_0(t) - \alpha p_0(t) \Delta t$$

Wir können diese Gleichung nun leicht zum Differentialquotienten umformen

$$\frac{p_0(t+\Delta t) - p_0(t)}{\Delta t} = -\alpha p_0(t)\Delta t$$

Indem wir $\Delta t \to 0$ gehen lassen, haben wir also die Differentialgleichung $\dot{p}_0(t) = -\alpha p_0(t)$ gefunden. Und wegen $p_0(0) = 1$ wird diese eindeutig gelöst:

$$p_0(t) = e^{-\alpha t}$$

Damit haben wir die Wahrscheinlichkeit gefunden, dass das Ereignis im Zeitintervall $[0, t]$ *nicht* eintritt. Sei nun allgemeiner $p_k(t)$ die Wahrscheinlichkeit für das k-malige Eintreten des Ereignisses im Zeitintervall $[0, t]$. Damit setzt sich $p_k(t + \Delta t)$ zusammen als Kombination: k Ereignisse bis t und danach kein Weiteres *oder* $k - 1$ Ereignisse bis t und danach noch ein Weiteres

$$p_k(t + \Delta t) = p_k(t) \cdot p_0(\Delta t) + p_{k-1}(t) \cdot p_1(\Delta t)$$

Setzen wir wieder $p_1(\Delta t) = \alpha\Delta t$ und $p_0(\Delta t) = 1 - \alpha\Delta t$ in diese Gleichung ein, so erhalten wir daraus dann weiter

$$p_k(t + \Delta t) = p_k(t) - \alpha p_k(t)\Delta t + \alpha p_{k-1}(t)\Delta t$$

Formen wir das wieder zu einem Differentialquotienten für p_k um und betrachten den Grenzwert $\Delta t \to 0$ so ergibt sich das DGL-System

$$\dot{p}_k(t) = -\alpha p_k(t) + \alpha p_{k-1}(t)$$

Da $p_0(t) = \exp(-\alpha t)$ schon bekannt ist, zeigt man mit Hilfe einer vollständigen Induktion über k leicht, dass daraus folgt

$$p_k(t) = \frac{(\alpha t)^k}{k!} e^{-\alpha t}$$

Insgesamt betrachten wir das Zeitintervall $[0, T]$ und setzen $\lambda := \alpha T$. Dann ist die Wahrscheinlichkeit für k Ereignisse also wie gesehen

$$p_k := p_k(T) = \frac{\lambda^k}{k!} e^{-\lambda}$$

Was genau ist die Bedeutung des Parameters λ? Treten im Schnitt k Ereignisse ein, dann ist $\Delta t = T/k$ also die durchschnittliche Zeit für das einmalige Vorkommen eines Ereignisses. Es stellt sich heraus, dass $\lambda = k$ gerade die mittlere Zahl von Ereignissen ist und damit sieht man also

$$\lambda = \frac{T}{\Delta t} = \frac{\text{betrachteter Zeitraum}}{\text{durchschnittliche Zeit für ein Ereignis}}$$

Definition 8.11: Poisson-Verteilung
Ist $\lambda \geq 0$ ein beliebiger positiver Parameter, dann definieren wir die **Poisson-Verteilung** $\mathcal{P}(\lambda)$ als den Laplace-Raum $(\mathbb{N}, p)$, wobei

$$p \ : \ \mathbb{N} \to [0,1] \ : \ k \mapsto \frac{\lambda^k}{k!} e^{-\lambda}$$

Satz 8.12:

(i) Ist $(\mathbb{N}, p) = \mathcal{P}(\lambda)$ die Poisson-Verteilung zum Parameter λ und bezeichnet $\mathrm{id} : \mathbb{N} \to \mathbb{N} \subseteq \mathbb{R} : k \mapsto k$ die Zufallsvariable der Zählung der Ereignisse, dann gilt

$$\begin{aligned} E(\mathrm{id}) &= \lambda \\ V(\mathrm{id}) &= \lambda \end{aligned}$$

(ii) Die Wahrscheinlichkeiten $p(k)$ der Poisson-Verteilung $(\mathbb{N}, p) = \mathcal{P}(\lambda)$ genügen einer einfachen Rekursionsformel (für alle $k \in \mathbb{N}$)

$$p(k+1) = \frac{\lambda}{k+1} p(k)$$

Mit Hilfe der Stirlingschen Formel kann man für große k aber auch die folgende Abschätzung verwenden:

$$p(k) \approx \frac{\exp\left(k + k \cdot \ln(\lambda/k) - \lambda\right)}{\sqrt{2\pi(k+1/6)}}$$

(iii) Ist $x = (x_1, \ldots, x_n) \in \mathbb{N}^n$ ein Datentupel (x_i ist die Anzahl der Vorkommens des Ereignisses im i-ten Versuch), dann ist die Maximum Likelihood Schätzung des Parameters λ das Tupel x unter id zu beobachten, gegeben durch das arithmetische Mittel

$$\textsc{mle}(\mathrm{id}, x) = A(x)$$

(iv) *Grenzwertsatz von Poisson*: Die Binomialverteilung kann für kleine Eintrittswahrscheinlichkeiten und viele Versuche durch die Poisson-Verteilung angenähert werden. Genauer gilt: Ist $\lambda > 0$ vorgelegt und ist $(p_n) \subseteq [0,1]$ eine Folge von Wahrscheinlichkeiten, so dass $(np_n) \to \lambda$ für $n \to \infty$ gilt, dann gilt im Grenzwert auch für alle $k \in \mathbb{N}$

$$\lim_{n\to\infty} b(n, p_n, k) = \frac{\lambda^k}{k!} e^{-\lambda}$$

(v) Ein Spezialfall des Grenzwertsatzes von Poisson ist folgende Faustformel: für $n \geq 11$, $p < 1/20$ und $k \in \mathbb{N}$, so dass $kp << 1$ und $k^2/n << 1$ hinreichend klein sind, kann man mit $\lambda := np$ abschätzen:

$$\binom{n}{k} p^k (1-p)^{n-k} = b(n,p,k) \approx p_\lambda(k) = \frac{\lambda^k}{k!} e^{-\lambda}$$

(vi) Zu jedem $i \in 1 \ldots m$ betrachten wir eine Poisson-Verteilung $(\mathbb{N}, p_i) = \mathcal{P}(\lambda_i)$ zum Parameter $\lambda_i \geq 0$ und bezeichnen mit $\mathrm{id}_i : \mathbb{N} \to \mathbb{R} : k_i \mapsto k_i$ wieder die Zufallsvariable der Zählung des i-ten Ereignisses. Sei weiterhin $(\mathbb{N}, p) = \mathcal{P}(\lambda_1 + \ldots + \lambda_m)$ die Poisson-Verteilung zur Summe der Parameter λ_i. Dann gilt für die Wahrscheinlichkeit über alle $\mathcal{P}(\lambda_i)$ insgesamt $k \in \mathbb{N}$ Vorkommen zu beobachten

$$\begin{aligned} p\left(\sum_{i=1}^{m} \mathrm{id}_i = k\right) &:= \sum_{|\alpha|=k} \prod_{i=1}^{m} p_i(\mathrm{id}_i = \alpha_i) \\ &= \frac{(\lambda_1 + \ldots + \lambda_m)^k}{k!} e^{-(\lambda_1 + \ldots + \lambda_m)} \quad = \quad p(k) \end{aligned}$$

Kapitel 9

Markov-Ketten

In diesem Abschnitt wenden wir uns folgendem Problem zu: gegeben ist eine endliche Anzahl von Zuständen 1 bis n. Wir betrachten nun ein Objekt, dass in diskreten Zeitabständen zwischen diesen Zuständen wechseln kann. Beispiel: eine Katze befindet sich entweder draußen, auf dem Sofa, unter dem Bett oder in der Küche. Wir sehen jede Stunde nach, wo sich die Katze befindet und notieren die Folge der Aufenthaltsorte. Nun interessieren wir uns für Fragen wie: mit welcher Wahrscheinlichkeit ist die Katze (auf lange Sicht gesehen) auf dem Sofa? Dieses Problem hat viele praktische Anwendungen: Migrationsbewegungen, Altersverteilungen, Wählerwanderung, und der Page Rank Algorithmus von Google, um nur einige zu nennen.

Übergangsmatrizen:
Wir betrachten die Menge der Zustände $S_0 := 1 \dots n$, zum Beispiel nummerierte Urnen. Zu Beginn befinden sich $a_i \in \mathbb{N}$ Objekte im Zustand $i \in S_0$, in Urnenbeispiel sind also a_i Kugeln in Urne i. Ist $A = a_1 + a_2 + \cdots + a_n$ die Gesamtzahl an Objekten/Kugeln, dann ist die Ausgangsverteilung $q_0 \in \mathbb{R}^n$ damit gegeben durch

$$q_0 \;=\; (q_0[1], \dots, q_0[n]) \quad \text{wobei} \quad q_0[i] \;=\; \frac{a_i}{A}$$

In jedem Schritt hat jedes Objekt eine gewisse Wahrscheinlichkeit von einem Zustand in einen anderen Zustand über zu gehen. Die Wahrscheinlichkeit soll dabei aber nicht von dem jeweiligen Objekt, sondern nur vom gegenwärtigen Zustand $j \in 1 \dots n$ und möglichem nächsten Zustand $i \in 1 \dots n$ abhängen. In k-ten Schritt (wobei $1 \leq k \in \mathbb{N}$) sei die Übergangswahrscheinlichkeit von Zustand j nach Zustand i gegeben, durch

$$P_k[i,j] \;\in\; [0,1]$$

Wie sieht dann die Verteilung q_1 der Kugeln nach einem Schritt aus? Es wird der Anteil $P_1[i,j]$ der Kugeln von Urne j nach Urne i übergegangen sein. Da Aq_j Kugeln in Urne j waren also $AP[i,j]q_j$. In Urne i sammeln sich aber Kugeln aus allen Urnen $j \in 1 \dots n$, so dass sich deren Anzahl summiert. Der Anteil $q_1[i]$ der Kugeln in Urne i nach einem Schritt ist also

$$q_1[i] \quad = \quad \sum_{j=1}^{n} P_1[i,j]\, q_0[j]$$

Dies gilt nun für jedes $i \in 1 \dots n$ und damit ist das n-Tupel $q_1 \in \mathbb{R}^n$ nichts anderes als das Ergebnis der Vektor-Matrix-Multiplikation $q_1 = P_1 q_0$ wobei

$$P_k \quad = \quad \begin{pmatrix} P_k[1,1] & P_k[1,2] & \dots & P_k[1,n] \\ P_k[2,1] & P_k[2,2] & & P_k[2,n] \\ \vdots & & \ddots & \vdots \\ P_k[n,1] & P_k[n,2] & \dots & P_k[n,n] \end{pmatrix}$$

Im nächsten Schritt verteilen sich die Kugeln neu - diesmal mit den Übergangswahrscheinlichkeiten P_2 so dass die Verteilung nach zwei Schritten gerade $q_2 = P_2 q_1 = P_2 P_1 q_0$ ist. Setzt man diese Überlegung fort, so ergibt sich für die Verteilung q_m nach m Schritten:

$$q_m \quad = \quad P_m \dots P_2 P_1 q_0$$

Klar ist, dass sich aus den Verteilungen q_m die Anzahl der Kugel in Urne $i \in 1 \dots n$ rekonstruieren lässt, als $A \cdot q_m[i]$. Damit genügt es sich auf die Verteilungen q_k zu beschränken und die Gesamtzahl A zu ignorieren.

Wir interessieren uns nun für die Entwicklung der Folge $(q_0, q_1, \dots, q_m)$ der Verteilungen und insbesondere für die Grenzverteilung q_m für $m \to \infty$. Es wird sich zeigen, dass für die Beantwortung dieser Fragen ein gutes Stück lineare Algebra nötig ist. Wir gehen zwar davon aus, dass die lineare Algebra als Basiswissen gut bekannt ist, wiederholen aber ein paar Notationen und ausgewählte Ergebnisse. Es gibt eine große Anzahl sehr guter Bücher und Skripten zu diesem Gebiet, von denen wir insbesondere [Friedberg, Insel, Spence; Linear Algebra] empfehlen wollen. Eine sehr lesenswerte Verallgemeinerung zur linearen Algebra über Ringen (genannt Modultheorie) findet sich in [Adkins, Weintraub; Algebra - an Approach via Module Theory].

Doch zunächst sollten wir betrachten, um was für Matrizen es sich bei den P_k überhaupt handelt: Zunächst ist q_0 eine Wahrscheinlichkeitsverteilung auf S_0 d.h. es ist $q_0[1] + \dots + q_0[n] = 1$. Ebenso sollen die Kugeln ja nur umverteilt werden. D.h. für alle $k \in \mathbb{N}$ und alle $j \in 1 \dots n$ muss die Summe aller aus Zustand j abgehenden Anteile $P_k[1,j] + \dots + P_k[n,j] = 1$ sein. Dies führt uns auf die Definition:

Definition 9.1: Stochastische Matrizen:
Wir nennen ein n-Tupel $p = (p_1, \ldots, p_n) \in \mathbb{R}^n$ reeller Zahlen **stochastisch**, falls alle Einträge positiv sind und deren Summe 1 ergibt, d.h. falls

(1) $\forall\, i \in 1 \ldots n$ gilt $p_i \geq 0$

(2) und $\sum_{i=1}^{n} p_i = 1$

In anderen Worten: p ist eine Wahrscheinlichkeitsverteilung auf $1 \ldots n$. Die Menge aller stochastischen n-Tupel wird als n-**Simplex** bezeichnet

$$\Delta_n \;:=\; \{\, p \in \mathbb{R}^n \mid p \text{ stochastisch} \,\}$$

Eine $(m \times n)$-Matrix $P \in \mathrm{mat}_{m,n}(\mathbb{R})$ wird **stochatisch** (oder genauer: *spalten-stoschastisch*) genannt, falls alle Spalten von P stochastische m-Tupel sind:

$$P \text{ stochastisch} \quad :\Longleftrightarrow \quad \forall\, j \in 1 \ldots n \;:\; \mathrm{col}_j(P) \in \Delta_m$$

Zunächst betten wir die oben ausgeführte Betrachtung (zu Übergangsmatrizen) in unsere bisherige Theorie der Laplace-Räume ein. Die dort vorgestellten Ergebnisse dürften wenig überraschen - sie entsprechen genau der Intuition die man vom Übergangsprozess hat. Um diese Art von Prozessen weiter zu untersuchen, werden wir aber zunächst Methoden der linearen Algebra aufbauen müssen.

Satz 9.2:
Seien $1 \leq m,\, n \in \mathbb{N}$ dann betrachten wir den Raum aller $(m+1)$-Tupel mit Einträgen aus $1 \ldots n$ und bezeichnen diesen, mit

$$S_m \;:=\; \{\, (s_0, s_1, \ldots, s_m) \mid \forall\, i \in 0 \ldots m : s_i \in 1 \ldots n \,\}$$

Ist nun $q \in \Delta_n$ ein stochastisches n-Tupel und ist für jedes $k \in 1 \ldots m$ eine quadratische, stochastische Matrix $P_k \in \mathrm{mat}_n(\mathbb{R})$ gegeben, dann erhalten wir einen Laplace-Raum (S_m, p_m) unter der Wahrscheinlichkeitsverteilung $p_m : S_m \to \mathbb{R}$ wobei

$$p_m(s_0, s_1, \ldots, s_m) \;:=\; P_m[s_m, s_{m-1}] \,\ldots\, P_1[s_1, s_0]\, q[s_0]$$

Ist nun $k \in 0 \ldots m$, dann definieren wir die Zufallsvariable ω_k auf S_m die in jeder Folge $s \in S_m$ den Zustand im k-ten Schritt auswertet

$$\omega_k \;:\; S_m \to 1 \ldots n \;:\; (s_0, s_1, \ldots, s_m) \mapsto s_k$$

Sind $a,\, b \in 1 \ldots n$ dann schreiben wir, wie in der Stochastik üblich, auch $\{\, \omega_k = a \,\}$ für die Menge $\omega_k^{-1}(a) = \{\, s \in S_m \mid \omega_k(s) = a \,\}$. Und für diese gelten dann die folgenden Aussagen:

(i) Sei $0 \leq k \leq m$ vorgelegt und zu jedem $i \in 0 \ldots k$ sei ein $a_i \in 1 \ldots n$ gegeben. Dann ist die Wahrscheinlichkeit die anfänglichen Zustände $(a_0, a_1, \ldots, a_k)$ zu beobachten unabhängig von m, denn es gilt:

$$p_m\left(\bigcap_{i=0}^{k} \{\, \omega_i = a_i \,\}\right) \quad = \quad P_k[a_k, a_{k-1}] \ldots P_1[a_1, a_0]\, q[a_0]$$

(ii) Und sind $a, b \in 1 \ldots n$, dann ist die Übergangswahrscheinlichkeit, dass im k-ten Schritt ($k \in 1 \ldots m$) Zustand a in Zustand b übergeht gerade

$$p_m(\omega_k = b \mid \omega_{k-1} = a) \quad = \quad P_k[b, a]$$

(iii) Sei $1 \leq k \leq m$ und $(a_0, a_1, \ldots, a_k)$ eine vorgelegte Folge von Zuständen $a_i \in 1 \ldots n$. Mit V bezeichnen wir die Menge aller $s \in S_m$ mit dieser *Vergangenheit*, also $V := \{\, \omega_0 = a_0 \,\} \cap \cdots \cap \{\, \omega_{k-1} = a_{k-1} \,\}$. Dann hängt die Wahrscheinlichkeit im k-ten Schritt im Zustand a_k anzukommen nur vom Zustand a_{k-1} ab, nicht von der ganzen Vergangenheit V. D.h. ist $p_m(V) \neq 0$ dann gilt

$$p_m(\omega_k = a_k \mid V) \quad = \quad p_m(\omega_k = a_k \mid \omega_{k-1} = a_{k-1})$$

Beweis:
Wir beweisen die Aussage durch Induktion über m: Der Fall $m = 1$ ist die Tatsache, dass $r := P_1 q \in \Delta_n$ wieder stochastisch ist, was leicht nachzurechnen ist und was wir später etwas allgemeiner zeigen werden. Im Induktionsschritt $m - 1 \to m$ können wir voraus setzen, dass jedes (S_{m-1}, p_{m-1}) ein Laplace-Raum ist. Dann rechnen wir nach, dass

$$\begin{aligned}
\sum_{s \in S_m} p_m(s) \quad &= \quad \sum_{s_0=1}^{n} \cdots \sum_{s_m=1}^{n} P_m[s_{m+1}, s_m] \ldots P_2[s_2, s_1]\, P_1[s_1, s_0]\, q[s_0] \\
&= \quad \sum_{s_1=1}^{n} \cdots \sum_{s_m=1}^{n} P_m[s_{m+1}, s_m] \ldots P_2[s_2, s_1] \sum_{s_0=1}^{n} P_1[s_1, s_0]\, q[s_0] \\
&= \quad \sum_{s_1=1}^{n} \cdots \sum_{s_m=1}^{n} P_m[s_{m+1}, s_m] \ldots P_2[s_2, s_1]\, r[s_1]
\end{aligned}$$

D.h. die Summe über alle $p_m(s)$ (wobei $s \in S_m$) mit Ausgangsverteilung q ist wiederum die Summe über $p_{m-1}(s)$ (wobei $s \in S_{m-1}$) mit Ausgangsverteilung r. Nach Induktionsvoraussetzung ist diese aber 1 und damit ist (S_m, p_m) selbst ein Laplace-Raum. Wir beweisen nun die Aussage (i) des Satzes

$$p_m\left(\bigcap_{i=0}^{k} \{\, \omega_i = a_i \,\}\right) \quad = \quad P_k[a_k, a_{k-1}] \ldots P_1[a_1, a_0]\, q[a_0]$$

Dazu bezeichnen wir $R_k := P_k[a_k, a_{k-1}] \dots P_1[a_1, a_0]q[a_0]$ und weiter auch $r[i] := P_{k+1}[i, a_k]$. D.h. $r = (r[1], \dots, r[n]) \in \Delta_n$ ist die a_k-te Spalte von P_{k+1} und damit ein stochastisches n-Tupel. Schließlich kürzen wir noch die oben genannte Teilmenge $V_k := \{\, \omega_0 = a_0 \,\} \cap \dots \cap \{\, \omega_k = a_k \,\} \subseteq S_m$ ab. Mit diesen Bezeichnungen werten wir dann diese Wahrscheinlichkeit aus

$$\begin{aligned}
p_m(V_k) &= \sum_{s_{k+1}=1}^{n} \cdots \sum_{s_m=1}^{n} p_m(a_0, a_1, \dots, a_k, s_{k+1}, \dots, s_m) \\
&= \sum_{s_{k+1}=1}^{n} \cdots \sum_{s_m=1}^{n} P_m[s_m, s_{m-1}] \dots P_{k+1}[s_{k+1}, a_k] \dots P_0[a_0] \\
&= \sum_{s_{k+1}=1}^{n} \cdots \sum_{s_m=1}^{n} P_m[s_m, s_{m-1}] \dots r[s_{k+1}] \cdot R_k \\
&= R_k \cdot \sum_{s_{k+1}=1}^{n} \cdots \sum_{s_m=1}^{n} P_m[s_m, s_{m-1}] \dots P_{k+2}[s_{k+2}, s_{k+1}] r[s_{k+1}]
\end{aligned}$$

Die Summe ist einfach die Gesamtwahrscheinlichkeit von (S_{m-k-1}, p_{m-k-1}) zur Ausgangsverteilung r und damit gleich 1. Daher ist $p_m(V_k) = R_k$ wie behauptet. Für Aussage (ii) betrachten wir der Einfachheit halber $m = k$ (im Fall $k < m$ treten nur weitere Faktoren auf, die aber in den folgenden beiden Ausdrücken gleich lauten). Zunächst werten wir $p_k(\omega_{k-1} = a)$ aus. Dies ist also der Ausdruck für die Gesamtwahrscheinlichkeit, in dem s_{k-1} durch a ersetzt wurde:

$$\begin{aligned}
&= \sum_{s_0=1}^{n} \cdots \sum_{s_{k-2}=1}^{n} \sum_{s_k=1}^{n} P_k[s_k, a] P_{k-1}[a, s_{k-2}] \dots P_1[s_1, s_0] q[s_0] \\
&= \sum_{s_k=1}^{n} P_k[s_k, a] \sum_{s_0=1}^{n} \cdots \sum_{s_{k-2}=1}^{n} P_{k-1}[a, s_{k-2}] \dots P_1[s_1, s_0] q[s_0] \\
&= 1 \cdot \sum_{s_0=1}^{n} \cdots \sum_{s_{k-2}=1}^{n} P_{k-1}[a, s_{k-2}] \dots P_1[s_1, s_0] q[s_0]
\end{aligned}$$

Letzteres, da die vordere Summe einfach über die a-te Spalte von P_k läuft und daher gleich 1 ist, da P_k stochastisch ist. Als zweiten Baustein betrachten wir $p_k(\omega_k = b$ und $\omega_{k-1} = a)$. Dies ist der Ausdruck für die Gesamtwahrscheinlichkeit in dem s_k durch b und s_{k-1} durch a ersetzt wurde. Er lautet:

$$\begin{aligned}
&= \sum_{s_0=1}^{n} \cdots \sum_{s_{k-2}=1}^{n} P_k[b, a] P_{k-1}[a, s_{k-2}] \dots P_1[s_1, s_0] q[s_0] \\
&= P_k[b, a] \sum_{s_0=1}^{n} \cdots \sum_{s_{k-2}=1}^{n} P_{k-1}[a, s_{k-2}] \dots P_1[s_1, s_0] q[s_0] \\
&= P_k[b, a] \cdot p_k(\omega_{k-1} = a)
\end{aligned}$$

Ist $p_k(\omega_{k-1} = a) \neq 0$ dann ist also die bedingte Wahrscheinlichkeit von $\omega_k = b$ unter der Bedingung $\omega_{k-1} = a$ definiert als der Quotient

$$p_m(\omega_k = b \mid \omega_{k-1} = a) \;=\; \frac{p_m(\omega_k = b \text{ und } \omega_{k-1} = a)}{p_m(\omega_{k-1} = a)} \;=\; P_k[b, a]$$

Nachdem nun (i) und (ii) gezeigt wurden, folgt die Behauptung (iii) sofort:

$$\begin{aligned} p_m(\omega_k = a_k \mid V_{k-1}) &= \frac{p_m(\{\,\omega_k = a_k\,\} \cap V_{k-1})}{p_m(V_{k-1})} = \frac{p_m(V_k)}{p_m(V_{k-1})} \\ &= \frac{P_k[a_k, a_{k-1}] \ldots P_1[a_1, a_0] q[a_0]}{P_{k-1}[a_{k-1}, a_{k-2}] \ldots P_1[a_1, a_0] q[a_0]} \\ &= P_k[a_k, a_{k-1}] \;=\; p_m(\omega_k = a_k \mid \omega_{k-1} = a_{k-1}) \end{aligned}$$

□

Bemerkung 9.3:
Die in dem Satz mit (iii) nummerierte Eigenschaft nimmt man als Definition für den Begriff einer Markov-Kette: Sei (S, p) ein Laplace-Raum und seien die $X_k : S \to \mathbb{R}$ Zufallsvariablen auf S (wobei $k \in 0 \ldots m$ oder $k \in \mathbb{N}$). Ist nun $1 \leq k$ und $a = (a_0, a_1, \ldots, a_k) \in \mathbb{R}^{k+1}$ dann bezeichnen wir

$$V(a) \;:=\; \bigcap_{i=0}^{k-1} \{\, X_i = a_i \,\}$$

Dann nennen wir die Folge von Zufallsvariablen (X_k) eine **Markov-Kette** auf S, falls für alle $1 \leq k$ und alle $(a_0, a_1, \ldots, a_k) \in \mathbb{R}^{k+1}$ gilt

$$p(V(a)) \;=\; 0 \quad \text{oder} \quad p(X_k = a_k \mid V(a)) \;=\; p(X_k = a_k \mid X_{k-1} = a_{k-1})$$

Anschaulich gesprochen bedeutet das, dass die Auswertung von X_k nur von dem bisherigen Zustand X_{k-1} abhängt, also kein Gedächtnis hat. Mit diesem Begriff hätte man (iii) also auch wie folgt formulieren können: ist $q \in \Delta_n$ und sind die $(P_k) \in \mathrm{mat}_n(\mathbb{R})$ stochastische Matrizen, dann ist die Folge (ω_k) der angenommenen Zustände eine Markov-Kette.

Komplexwertige Vektoren:
Wie der $\mathbb{R}^m$ eine Algebra über dem Grundkörper $\mathbb{R}$ ist, ist $\mathbb{C}^m$ eine Algebra über dem Grundkörper $\mathbb{C}$ - vergleiche dazu 2.2. Die Verknüpfungen der Addition, Multiplikation und Skalarmultiplikation lassen sich wörtlich übertragen, lediglich $\mathbb{R}$ muss jeweils durch $\mathbb{C}$ ersetzt werden. Das *Skalarprodukt* gibt es jetzt aber in 2 Fassungen: einmal bilinear und einmal sesquilinear (d.h. linear im ersten (w) und konjugiert-linear im zweiten (z) Argument). Seien $z = (z_1, \ldots, z_m)$ und $w = (w_1, \ldots, w_m) \in \mathbb{C}^m$ dann setzen wir:

$$\begin{aligned} w \bullet z &:= w_1 z_1 + w_2 z_2 + \cdots + w_m z_m \\ w \circ z &:= w_1 \overline{z}_1 + w_2 \overline{z}_2 + \cdots + w_m \overline{z}_m \\ \|z\| &:= \sqrt{z \circ z} \;=\; \sum_{i=1}^{m} |z_i|^2 \end{aligned}$$

Die *natürliche Norm* $\|z\|$ ist aufgrund der Konstruktion von $z \circ z$ wieder eine positive, reelle Zahl. Sie ist aber nicht die einzige interessante Norm auf $\mathbb{C}^m$. Ganz analog kann man für jedes $1 \leq p \in \mathbb{N}$ und sogar für $p = \infty$ weitere Normen auf $\mathbb{C}^m$ einführen:

$$\begin{aligned} \|z\|_1 &:= \sum_{i=1}^{m} |z_i| \\ \|z\|_p &:= \left(\sum_{i=1}^{m} |z_i|^p \right)^{\frac{1}{p}} \\ \|z\|_\infty &:= \max\{\, |z_i| \mid i \in 1 \ldots m \,\} \end{aligned}$$

Man beachte, dass $\|z\|_2 = \|z\|$ gerade wieder die natürliche Norm auf $\mathbb{C}^m$ ist. Und $\|z\|_\infty$ wird auch die *Supremumsnorm* genannt. Allen Normen gemeinsam sind die folgenden Eigenschaften für alle $a \in \mathbb{C}$ und z, $w \in \mathbb{C}^m$

$$\begin{aligned} \|az\| &= |a| \cdot \|z\| \\ \|w + z\| &\leq \|w\| + \|z\| \\ \|z\| = 0 &\iff z = 0 \end{aligned}$$

Alle Normen auf endlich-dimensionalen Vektorräumen (über lokal-kompakten Körpern) sind aber äquivalent. In unserem Fall gelten für $1 \leq p \leq q \leq \infty$ beispielsweise stets die folgenden Abschätzungen

$$\|z\|_q \;\leq\; \|z\|_p \;\leq\; m^{\frac{1}{p} - \frac{1}{q}} \, \|z\|_q$$

(wobei wir der Konsistenz halber $1/\infty = 0$ gesetzt haben). Und das bedeutet, dass alle Normen denselben Konvergenzbegriff definieren. D.h. konvergiert eine Folge $(z_n) \subseteq \mathbb{C}^m$ in der einen Norm, dann konvergiert sie bereits bezüglich aller Normen auf $\mathbb{C}^m$. Insbesondere gilt

$$\begin{aligned} (z_n) \to z \quad &:\iff \quad (\|z_n - z\|_p) \to 0 \\ &\iff \quad \forall\, i \in 1 \ldots m \;:\; (z_{i,n}) \to z_i \end{aligned}$$

Matrixalgebra:
Lineare Abbildungen $\mathbb{C}^n \to \mathbb{C}^m$ können durch $(m \times n)$-Matrizen repräsentiert werden. Eine Matrix $A \in \mathrm{mat}_{m,n}(\mathbb{C}) = (\mathbb{C}^m)^n$ notieren wir dabei in der Form

$$A \;=\; (a_{i,j}) \;=\; \begin{pmatrix} a_{1,1} & a_{1,2} & \dots & a_{1,n} \\ a_{2,1} & a_{2,2} & \dots & a_{2,n} \\ \vdots & \vdots & & \vdots \\ a_{m,1} & a_{m,2} & \dots & a_{m,n} \end{pmatrix}$$

Die i-te Reihe (wobei $i \in 1 \dots m$) von A bezeichnen wir mit $\mathrm{row}_i(A)$, die j-te Spalte (wobei $j \in 1 \dots n$) von A mit $\mathrm{col}_j(A)$. Und den Eintrag von A in der i-ten Reihe und j-ten Splalte bezeichnen wir mit $A[i,j]$

$$\begin{aligned} \mathrm{row}_i(A) &:= (a_{i,1}, a_{i,2}, \dots, a_{i,n}) \;\in\; \mathbb{C}^n \\ \mathrm{col}_j(A) &:= (a_{1,j}, a_{2,j}, \dots, a_{m,j}) \;\in\; \mathbb{C}^m \\ A[i,j] &:= a_{i,j} \;\in\; \mathbb{C} \end{aligned}$$

Die Matrix A wird dabei zur linearen Abbildung $A : \mathbb{C}^n \to \mathbb{C}^m$ durch Matrix-Vektor-Multiplikation $z \mapsto Az$. Und mit Hilfe der gerade eingeführten Notation, lässt sich diese schreiben, als

$$Az \;:=\; \begin{pmatrix} \mathrm{row}_1(A) \bullet z \\ \mathrm{row}_2(A) \bullet z \\ \vdots \\ \mathrm{row}_m(A) \bullet z \end{pmatrix} \;\in\; \mathbb{C}^m$$

Ist nun $B \in \mathrm{mat}_{\ell,m}(\mathbb{C})$ eine weitere Matrix, dann lässt sich auch die Matrix-Multiplikation $BA \in \mathrm{mat}_{\ell,n}(\mathbb{C})$ notieren, als

$$BA \;=\; (\mathrm{row}_i(B) \bullet \mathrm{col}_k(A)) \;=\; \begin{pmatrix} \vdots & & \vdots \\ B\,\mathrm{col}_1(A) & \cdots & B\,\mathrm{col}_n(A) \\ \vdots & & \vdots \end{pmatrix}$$

Natürlich lässt sich $\mathrm{mat}_{m,n}(\mathbb{C})$ mit $\mathbb{C}^{m \times n}$ identifizieren (es handelt sich um einen Vektorraum-Isomorphismus, die Multiplikation ist eine andere). Und damit kann man die p-Normen (für $1 \le p \le \infty$) auf die Matrizen übertragen:

$$\|A\|_p \;:=\; \left(\sum_{i=1}^{m} \sum_{j=1}^{n} |A[i,j]|^p \right)^{\frac{1}{p}}$$

Interessanter ist jedoch die folgende Norm, die auch also *Operatornorm* oder *Spektralnorm* auf $\mathrm{mat}_{m,n}(\mathbb{C})$ bezeichnet wird:

$$\| A \|_{\mathrm{op}} \;:=\; \sup\{\, \|Az\| \mid z \in \mathbb{C}^n \text{ mit } \|z\| = 1 \,\}$$

Man hätte dies natürlich auch bezüglich irgend einer anderen p-Norm definieren können, aber das macht keinen wesentlichen Unterschied. Wichtiger ist, dass die Operatornorm nicht nur eine Norm ist, sondern nach Konstruktion auch *submultiplikativ* ist: D.h. es gilt für alle $z \in \mathbb{C}^n$ und alle $B \in \mathrm{mat}_{\ell,m}(\mathbb{C})$

$$\begin{aligned} \|Az\| &\leq \| A \|_{\mathrm{op}} \cdot \|z\| \\ \| BA \|_{\mathrm{op}} &\leq \| B \|_{\mathrm{op}} \cdot \| A \|_{\mathrm{op}} \end{aligned}$$

Der Ehrlichkeit halber sollte man auch sagen, dass es weitere submultiplikative Normen gibt, wie etwa die 2-Norm $\|A\|_2$ oder $\sqrt{m \times n} \cdot \|A\|_\infty$. Ein umfassendes Skriptum zu diesem Thema findet man online bei [Jan Falkenhain, Operatornormen]. Wiederum ist $\mathrm{mat}_{m,n}(\mathbb{C})$ endlich-dimensional (Dimension $m \cdot n$) so dass alle Normen äquivalent sind. Für die Operatornorm findet man sofort die folgenden Abschätzungen:

$$\|A\|_\infty \leq \| A \|_{\mathrm{op}} \leq \|A\|_1$$

Wie alle Normen, induziert auch die Operatornorm einen Konvergenzbegriff auf dem Raum aller Matrizen $\mathrm{mat}_{m,n}(\mathbb{C})$. Ist (A_k) eine Folge und A eine $(m \times n)$-Matrix, dann

$$\begin{aligned} (A_k) \to A \quad &:\Longleftrightarrow \quad (\| A_k - A \|_{\mathrm{op}}) \to 0 \\ &\Longleftrightarrow \quad \forall\, i \in 1 \ldots m, \forall\, j \in 1 \ldots n \ : \ (A_k[i,j]) \to A[i,j] \end{aligned}$$

Satz 9.4:
Ist A eine einzelne und $(A_k) \subseteq \mathrm{mat}_{m,n}(\mathbb{C})$ eine Folge von $(m \times n)$-Matrizen, dann gelten die folgenden Aussagen:

(i) Konvergiert $(A_k) \to A$ und sind $L \in \mathrm{mat}_{\ell,m}(\mathbb{C})$ und $R \in \mathrm{mat}_{n,r}(\mathbb{C})$ zwei weitere, passende Matrizen, dann konvergiert auch die Folge

$$(LA_kR) \to LAR$$

(ii) Sind $S \in \mathrm{gl}_m(\mathbb{C})$ und $T \in \mathrm{gl}_n(\mathbb{C})$ quadratische, invertierbare Matrizen, dann sind die folgenden beiden Aussagen sogar äquivalent

(a) $(A_k) \to A$

(b) $(SA_kT) \to SAT$

Beweis:
Wegen der Submultiplikativität ist (i) klar: $LA_kR - LAR = L(A_k - A)R$ und damit $\| LA_kR - LAR \|_{\mathrm{op}} \leq \| L \|_{\mathrm{op}} \cdot \| A_k - A \|_{\mathrm{op}} \cdot \| R \|_{\mathrm{op}} \to 0$, da nach Vorausetzung ja $(A_k) \to A$ konvergiert. Damit folgt insbesondere die Implikation (a) $\Longrightarrow$ (b) in (ii). Und die Umkehrung (b) $\Longrightarrow$ (a) folgt dann ebenso, da $(A_k) = (S^{-1}(SA_kT)T^{-1}) \to S^{-1}(SAT)T^{-1} = A$.

$\square$

Eigenwerte:
Ist $A \in \mathrm{mat}_n(\mathbb{C})$ eine quadratische Matrix, dann betrachten wir eine Gleichung der Form (wobei $\lambda \in \mathbb{C}$ und $z \in \mathbb{C}^n$

$$Az \;=\; \lambda z$$

Diese Gleichung ist offensichtlich äquivalent zu $(A - \lambda E) = 0$, wobei E die Einheitsmatrix (also das Einselment von $\mathrm{mat}_n(\mathbb{C})$) bezeichnet. Eine Lösung besteht also offensichtlich immer: $z = 0$. Etwas Neues ergibt sich also nur, wenn man nach nicht-trivialen Lösungen $z \neq 0$ sucht. In diesem Fall ist $A - \lambda E$ also nicht-injektiv und muss daher Determinante 0 haben

$$c_A(\lambda) \;:=\; \det(A - \lambda E) \;=;\; 0$$

Dabei nennen wir $c_A(t) := \det(A - tE)$ das *charakteristische Polynom* von A. Offensichtlich hat $c_A(t)$ den Grad n und man sieht schnell, dass gilt

$$c_A(t) \;=\; (-1)^n t^n + (-1)^{n-1}\mathrm{tr}(A)t^{n-1} + \cdots + \det(A)$$

Wie wir gerade gesehen haben, hat die Gleichung $Az = \lambda z$ genau dann eine nicht-triviale Lösung $z \neq 0$, falls λ eine Nullstelle davon ist. Die Nullstellen von $c_A(t)$ werden daher *Eigenwerte* von A genannt. Die Menge aller Eigenwerte von A heißt das *Spektrum* von A

$$\mathrm{Spec}(A) \;:=\; \{\, \lambda \in \mathbb{C} \mid c_A(\lambda) = 0 \,\}$$

Die Lage des Spektrums von A ist aber nicht beliebig: zunächst gilt für alle $\lambda \in \mathrm{Spec}(A)$, dass $|\lambda| \leq \| A \|_{\mathrm{op}}$ ist. Und weiter gilt der Satz von Gerschgorin [Friedberg, Insel, Spence; Linear Algebra; 5.21] nach dem jeder Eigenwert von A in einer Kreisscheibe von Radius $r_i = \sum_{i \neq j} |A[i,j]|$ um $A[i,i]$ liegt:

$$\mathrm{Spec}(A) \;\subseteq\; \bigcup_{i=1}^{n} \{\, z \in \mathbb{C} \mid |z - A[i,i]| \leq r_i \,\}$$

Da $\mathbb{C}$ algebraisch abgeschlossen ist - das ist überhaupt der Grund warum wir von $\mathbb{R}$ auf $\mathbb{C}$ erweitert haben - zerfällt das charakteristische Polynom

$$c_A(t) \;=\; (-1)^n \prod_{\lambda \in \mathrm{spec}(A)} (t - \lambda)^{m(\lambda)}$$

Die dabei auftretende Potenz $m(\lambda)$ wird *algebraische Vielfachheit* des Eigenwertes λ genannt. Wie bezeichnen sie mit

$$\mathrm{algm}(A, \mu) \;:=\; \max\Big\{\, k \in \mathbb{N} \mid (t - \lambda)^k \;\mid\; c_A(t) \,\Big\} \;\in\; \mathbb{N}$$

Ist $z \in \mathbb{C}^n$ ein Tupel, mit $Az = \lambda z$, dann nennen wir z einen *Eigenvektor* von A zum Eigenwert λ. Die Menge aller Eigenvektoren von A zu λ wird als *Eigenraum* von A zu λ bezeichnet und ist offenbar ein Unterraum von $\mathbb{C}^n$

$$\mathrm{Eig}(A, \lambda) \;:=\; \mathrm{kn}(A - \lambda E) \;=\; \{\, z \in \mathbb{C}^n \mid Az = \lambda z \,\}$$

Die Dimension dieses Unterraumes wird als *geometrische Vielfachheit* bezeichnet, sie kann die algebraische Vielfachheit nie übersteigen

$$\mathrm{geom}(A,\mu) \;:=\; \dim(\mathrm{Eig}(A,\lambda)) \;\leq\; \mathrm{algm}(A,\mu)$$

Ein Beweis findet sich in [Friedberg, Insel, Spence; Linear Algebra; 5.12]. Es kann jedoch sein, dass mindestens ein Eigenraum zu klein ausfällt (d.h. dass die geometrische echt kleiner ist, als die algebraische Vielfachheit). In diesem Fall ist eine hat A keine Diagonalform, sondern nur eine Jordan-Normalform.

Wir wollen hier aber nur auf den Fall der Diagonalisierbarkeit einghen: Gibt es genügend viele Eigenvektoren, dann kann man eine Basis aus diesen bilden und offensichtlich nimmt A in der Basis ihrer Eigenvektoren Diagonalform an. Genauer gesagt sind die folgenden beiden Aussagen äquivalent

(a) $\forall\,\lambda \in \mathrm{Spec}(A)$ gilt $\mathrm{geom}(A,\lambda) = \mathrm{algm}(A,\lambda)$

(b) A ist *diagonalisierbar*: d.h. es gibt eine invertierbare Matrix $T \in \mathrm{gl}_n(\mathbb{C})$ und Zahlen $\lambda_1,\ldots,\lambda_n$ so dass gilt

$$T^{-1}AT \;=\; \begin{pmatrix} \lambda_1 & 0 & & 0 \\ 0 & \lambda_2 & & \\ & & \ddots & \\ 0 & & & \lambda_n \end{pmatrix}$$

In diesem Fall sind die λ_i natürlich die (nicht notwendig verschiedenen) Eigenwerte von A, d.h. es ist $\mathrm{Spec}(A) = \{\,\lambda_1,\ldots,\lambda_n\,\}$. Und die i-te Spalte $z_i = \mathrm{col}_i(T)$ von T ist ein Eigenvektor von A zum Eigenwert λ_i. Man beachte, dass dies gar nicht so selten ist - nach [Friedberg, Insel, Spence; Linear Algebra; 6.17 und 6.18] sind weiterhin äquivalent:

(a) A ist *normal*, d.h. bezeichnet $A^\dagger$ die transponierte und konjugierte Matrix von A, dann gilt $A^\dagger A = AA^\dagger$

(b) A ist *unitär diagonalisierbar*: d.h. es gibt eine invertierbare Matrix $T \in \mathrm{gl}_n(\mathbb{C})$ mit $T^{-1} = T^\dagger$ und Zahlen $\lambda_1,\ldots,\lambda_n$ so dass gilt

$$T^{-1}AT \;=\; \begin{pmatrix} \lambda_1 & 0 & & 0 \\ 0 & \lambda_2 & & \\ & & \ddots & \\ 0 & & & \lambda_n \end{pmatrix}$$

Diagonalisierung von Matrizen ist insbesondere hilfreich um Matrixpotenzen auszuwerten. Ist $T^{-1}AT = \Lambda$ die oben genannte Diagonalmatrix, dann gilt offenbar $T^{-1}A^kT = T^{-1}A(TT^{-1})A(T\ldots T^{-1})AT = (T^{-1}AT)^k = \Lambda^k$ also

$$A^k \;=\; T\Lambda^kT^{-1} \;=\; T\begin{pmatrix} \lambda_1^k & 0 & & 0 \\ 0 & \lambda_2^k & & \\ & & \ddots & \\ 0 & & & \lambda_n^k \end{pmatrix}T^{-1}$$

Allgemeiner kann man A damit in ein beliebiges Polynom $f \in \mathbb{C}[t]$ einsetzen (durch Reihenentwicklung sogar in eine beliebige holomorphe, auf ganz $\mathbb{C}$ definierte Funktion $f \in \mathcal{O}(\mathbb{C})$) und für dieses f gilt dann:

$$f(A) \;=\; T \begin{pmatrix} f(\lambda_1) & 0 & & 0 \\ 0 & f(\lambda_2) & & \\ & & \ddots & \\ 0 & & & f(\lambda_n) \end{pmatrix} T^{-1}$$

An dieser Darstellung erkennt man auch, dass die Folge (A^k) der Matrixpotenzen genau dann konvergiert, wenn die Folgen der Potenzen der Eigenwerte (λ_i^k) konvergieren. Und dies ist offenbar der Fall, wenn $|\lambda_i| < 1$ ist (in diesem Fall gilt $(\lambda_i^k) \to 0$), oder wenn $\lambda_i = 1$ ist (in diesem Fall ist $(\lambda_i^k) = 1$).

Doch diese Aussage lässt sich nach [Friedberg, Insel, Spence; Linear Algebra; 5.18] wie folgt auf nicht-diagonalisierbare Matrizen verallgemeinern: Für eine beliebige quadratische Matrix $A \in \mathrm{mat}_n(\mathbb{C})$ sind die folgenden beiden Aussagen äquivalent:

(a) Die Folge (A^k) konvergiert, d.h. es gibt eine Matrix $L \in \mathrm{mat}_n(\mathbb{C})$ so dass $(A^k) \to L$ gilt

(b) Es ist $\mathrm{Spec}(A) \subseteq \{\, z \in \mathbb{C} \mid |z| < 1 \,\} \cup \{\, 1 \,\}$ und ist $1 \in \mathrm{Spec}(A)$ dann gilt auch $\mathrm{algm}(A, 1) = \mathrm{geom}(A, 1)$.

Beispiel 9.5:
Wir betrachten den Fall $n = 2$, d.h. eine quadratische (2×2)-Matrix A. Wie üblich bezeichnen wir deren Koeffizienten mit a, b, c und d

$$A \;=\; \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

Das charakteristische Polynom lautet dann $c_A(t) = t^2 - (a+d)t + (ad - bc)$ und man erkennt sofort die Spur, respektive Determinante von A darin. Die quadratische Gleichung $c_A(\lambda) = 0$ lässt sich bekanntlich lösen durch

$$\lambda_{1,2} \;:=\; \frac{1}{2}\,(a + d \pm \delta) \quad \text{wobei} \quad \delta \;:=\; \sqrt{(a-d)^2 + 4bc}$$

Dies sind also die Eigenwerte von A. Die zugehörigen Eigenvektoren findet man durch Lösung des entsprechenden Gleichungssystems $Ae_i = \lambda_i e_i$ zu

$$e_1 \;=\; \begin{pmatrix} a - d + \delta \\ 2c \end{pmatrix} \quad \text{und} \quad e_2 \;=\; \begin{pmatrix} d - a + \delta \\ -2c \end{pmatrix}$$

Die Transformationsmatrix T ist also gegeben, durch $T = (e_1\ e_2)$, wobei die e_i in den beiden Spalten von T stehen. Damit nimmt $T^{-1}AT = \Lambda$ Diagonalgestalt an und wir können die Matrixpotenzen von A explizit berechnen

$$A^k \;=\; \frac{1}{2\delta} \begin{pmatrix} (d-a+\delta)\lambda_2^k + (a-d+\delta)\lambda_1^k & 2b(\lambda_1^k - \lambda_2^k) \\ 2c(\lambda_1^k - \lambda_2^k) & (a-d+\delta)\lambda_2^k + (d-a+\delta)\lambda_1^k \end{pmatrix}$$

Satz 9.6:

(i) Sei $P \in \mathrm{mat}_n(\mathbb{R})$ eine reelle, quadratische Matrix mit positiven Einträgen $P[i,j] \geq 0$ und bezeichne $e := (1,1,\ldots,1) \in \mathbb{R}^n$ dann gilt

$$P \text{ stochastisch} \quad \Longleftrightarrow \quad P^* e \;=\; e$$

wobei P^* die transponierte Matrix von P bezeichnet. Ist P stochastisch, dann ist also insbesondere $1 \in \mathrm{Spec}(P)$ ein Eigenwert von P.

(ii) Die Menge der stochastischen Matrizen ist konvex, d.h. sind P und $Q \in \mathrm{mat}_{m,n}(\mathbb{R})$ stochastische Matrizen und ist $\alpha \in [0,1]$ dann ist auch die Matrix $\alpha P + (1-\alpha)Q \in \mathrm{mat}_{m,n}(\mathbb{R})$ stochastisch.

(iii) Ist $P \in \mathrm{mat}_{m,n}(\mathbb{R})$ eine stochastische Matrix und $q \in \Delta_n$ ein stochastisches n-Tupel, dann ist $Pq \in \Delta_m$ wieder stochastisch.

(iv) Sind $P \in \mathrm{mat}_{m,n}(\mathbb{R})$ und $Q \in \mathrm{mat}_{\ell,m}(\mathbb{R})$ stochastische Matrizen, dann ist auch deren Produkt $QP \in \mathrm{mat}_{\ell,n}(\mathbb{R})$ stochastisch.

(v) Ist $P \in \mathrm{mat}_n(\mathbb{R})$ eine quadratische, stochastische Matrix und ist $k \in \mathbb{N}$, dann ist auch die Potenz $P^k \in \mathrm{mat}_n(\mathbb{R})$ stochastisch.

(vi) Ist $P \in \mathrm{mat}_n(\mathbb{R})$ eine stochastische Matrix, dann haben alle Eigenwerte von P höchstens den Betrag 1, d.h. es gilt

$$\mathrm{Spec}(P) \quad \subseteq \quad \{\, z \in \mathbb{C} \mid |z| \leq 1 \,\}$$

Beweis:

(i) Die j-te Zeile von P^* ist nach Konstruktion die j-te Spalte von P, d.h. für $j \in 1 \ldots n$ gilt: $\mathrm{row}_j(P^*) = \mathrm{col}_j(P)$. Bezeichnen wir $q := P^* e$, dann ist der j-te Koeffizient von q also $q_j = \mathrm{col}_j(P) \bullet e = \sum_i P[i,j] \cdot 1 = \sum_i P[i,j]$. Also ist $q = e$ äquivalent zu $\sum_i P[i,j] = 1$ für alle $j \in 1 \ldots n$. Und das war gerade die Definition von P stochastisch zu sein. Ist nun P stochastisch, dann ist damit e ein Eigenvektor von P^* zum Eigenwert 1 und daher $1 \in \mathrm{Spec}(P^*)$. Bezeichnen wir $B := P - \lambda E$, dann ist $B^* = P^* - \lambda E$. Die Determinante ist jedoch invariant $\det(B) = \det(B^*)$ unter der Transposition von Matrizen. Also haben die charakteristischen Polynome von P und P^* dieselben Nullstellen und damit gilt auch $1 \in \mathrm{Spec}(P)$.

(ii) Die Aussage ist unmittelbar einsichtig: sei $R := \alpha P + (1-\alpha)Q$, also $R[i,j] = \alpha P[i,j] + (1-\alpha)Q[i,j]$. Wegen $\alpha \geq 0$ und $1-\alpha \geq 0$ sind damit auch alle Einträge $R[i,j]$ von R positiv. Und die j-te Spalte von R gegeben, durch $\mathrm{col}_j(R) = \alpha \mathrm{col}_j(P) + (1-\alpha)\mathrm{col}_j(Q) \in \Delta_m$. Die Spaltensumme ist wieder gleich 1, da $\alpha \cdot 1 + (1-\alpha) \cdot 1 = 1$.

(iii) Sei $P = (p_{i,j})$ wobei $i \in 1 \ldots m$ und $j \in 1 \ldots n$ uns bezeichne $r = Pq$ also $r_i = \sum_j p_{i,j} q_j \geq 0$. Dann rechnet man nach, dass

$$\sum_{i=1}^{m} r_i \;=\; \sum_{i=1}^{m} \sum_{j=1}^{n} p_{i,j} q_j \;=\; \sum_{j=1}^{n} q_j \sum_{i=1}^{m} p_{i,j} \;=\; \sum_{j=1}^{n} q_j \;=\; 1$$

(iv) Nach Definition der Matrixmultiplikation gilt $\mathrm{col}_j(QP) = Q\,\mathrm{col}_j(P)$ und da nach Voraussetzung $\mathrm{col}_j(P) \in \Delta_m$ ist, folgt aus (i) unmittelbar auch $\mathrm{col}_j(QP) \in \Delta_\ell$. D.h. QP ist ebenfalls stochastisch.

(v) Für $k = 0$ ist $P^k = E$ offensichtlich stochastisch. Und der Rest ist klar mit Induktion: ist P^k stochastisch, dann nach (ii) auch $P^{k+1} = PP^k$.

(vi) Ist $\lambda \in \mathrm{Spec}(P)$ ein beliebiger Eigenwert von P, dann gibt es nach dem Satz von Gerschgorin ein $i \in 1 \ldots n$, so dass $|\lambda - P[i,i]| \leq r_i = \sum_{j \neq i} P[i,j]$ (die Beträge können wir uns sparen, da ja alle $P[i,j]$ positiv sind). Nun gilt aber die Dreiecksungleichung nach unten:

$$|\lambda| - P[i,i] \;\leq\; \sum_{j \neq i} P[i,j] \quad \Longrightarrow \quad |\lambda| \;\leq\; \sum_{j=1}^{n} P[i,j] \;=\; 1$$

□

Satz 9.7:
[Friedberg, Insel, Spence; Linear Algebra; 5.25] bzw. [Renyi, Wahrscheinlichkeitsrechnung; VEB 1979] Sei $P \in \mathrm{mat}_n(\mathbb{R})$ eine quadratische, stochastische Matrix. Wir nehmen an, es gibt eine Potenz P^m von P so dass alle Einträge in einer Zeile von P^m echt positiv sind, formal

$$\exists\, m \in \mathbb{N} \;\; \exists\, i \in 1 \ldots n \;\; : \;\; \forall\, j \in 1 \ldots n \;\; : \;\; P^m[i,j] \;>\; 0$$

(i) Dann ist $1 \in \mathrm{Spec}(P) \subseteq \{\, z \in \mathbb{C} \mid |z| < 1 \,\} \cup \{\, 1 \,\}$ und die algebraische und geometrische Vielfachheit des Eigenwertes 1 ist 1

$$\mathrm{algm}(P,1) \;=\; \mathrm{algm}(P,1) \;=\; 1$$

(ii) Insbesondere konvergiert die Folge der Matrixpotenzen (P^k) und wir bezeichnen deren Grenzwert im folgenden mit $L \in \mathrm{mat}_n(\mathbb{R})$

$$L \;:=\; \lim_{k \to \infty} P^k$$

(iii) Damit ist L eine stochastische Matrix, für die $PL = LP = L$ gilt. Alle Spalten von L sind identisch, d.h. es gibt ein $\ell \in \Delta_n$ so dass

$$L \;=\; \begin{pmatrix} \vdots & & \vdots \\ \ell & \cdots & \ell \\ \vdots & & \vdots \end{pmatrix}$$

(iv) Dabei ist $\ell \in \Delta_n$ das eindeutig bestimmte stochastische n-Tupel, für das $P\ell = \ell$ gilt. Wir nennen ℓ die *stationäre Verteilung* oder auch die *Grenzverteilung* von P.

(v) *Ergodensatz*: Ist $q \in \Delta_n$ ein beliebiges stochastisches n-Tupel, dann ist der Grenzwert $(P^k q)$ unabhängig von der Ausgabgsverteilung q

$$\lim_{k \to \infty} P^k q \;=\; \ell$$

Beispiel 9.8:
Im Fall dass eine Potenz P^m nur echt positive Einträge hat, ist also alles schön. Ist dies nicht der Fall, dann können aber Komplikationen auftreten: Den Fall $n = 2$ haben wir bereits für beliebige Matrizen behandelt. Wir betrachten nun eine stochastische (3×3) Matrix $P \in \mathrm{mat}_3(\mathbb{R})$. Bezeichnet $s := \mathrm{tr}(P)$ die Spur und $d := \det(P)$ die Determinante von P, dann ist das charakteristische Polynom von P gegeben durch

$$c_P(t) \quad = \quad (1-t)(t^2 + (1-s)t + d)$$

Wie wir wissen, gilt für alle Eigenwerte $|\lambda| \leq 1$. Im allgemeinen kann P aber weitere Eigenwerte vom Betrag 1 besitzen (es ist sogar möglich, dass $\mathrm{algm}(P, 1) \geq 2$ und P nicht diagonalisierbar ist). Beispielsweise hat die folgende Matrix das charakteristische Polynom $(1-t)(t^2 + 0.5t + 0.5)$ und damit die Eigenwerte $\lambda_1 = 1$, $\lambda_2 = -1$ und $\lambda_3 = 0.5$

$$P \quad := \quad \begin{pmatrix} 0.5 & 0 & 0 \\ 0.5 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix} \quad \text{bzw.} \quad \Lambda \quad := \quad \begin{pmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 0.5 \end{pmatrix}$$

Die zugehörigen Eigenvektoren von P sind $e_1 = (0, 1, 1)$, $e_2 = (0, 1, -1)$ und $e_3 = (-3, 1, 2)$ so dass wir die folgende Transformationsmatrix erhalten:

$$T \quad = \quad \begin{pmatrix} 0 & 0 & -3 \\ 1 & 1 & 1 \\ 1 & -1 & 2 \end{pmatrix} \quad \text{bzw.} \quad T^{-1} \quad = \quad \frac{1}{6} \begin{pmatrix} 3 & 3 & 3 \\ -1 & 3 & -3 \\ -2 & 0 & 0 \end{pmatrix}$$

Und mit Hilfe der Formel $P^k = T\Lambda^k T^{-1}$ kann man daraus wieder die Matrixpotenz P^k berechnen. Man beachte, dass P^k nicht konvergiert, da der Eigenwert $\lambda_2 = -1$ die Konvergenz zerstört

$$P^k \quad = \quad \frac{1}{6} \begin{pmatrix} 6/2^k & 0 & 0 \\ 3 - (-1)^k - 2/2^k & 6(1 + (-1)^k) & 6(1 - (-1)^k) \\ 3 + (-1)^k - 4/2^k & 6(1 - (-1)^k) & 6(1 + (-1)^k) \end{pmatrix}$$

Beispiel 9.9:
Nachdem die Theorie nun etabliert ist, betrachten wir unser erstes, klassisches Beispiel eines Übergangsprozesses: wir ordnen die Zustände 1 bis 4 als Ecken eines Quadrates an. Auf diesem Quadrat krabbelt eine Spinne, die in 1 beginnt. In jedem Schritt hat die Spinne eine Wahrscheinlichkeit von p sitzen zu bleiben und eine Wahrscheinlichkeit von q eine Kante im Uhrzeigersinn weiter zu krabbeln. Die Wahrscheinlichkeit, dass die Spinne eine Kante gegen den Uhrzeigersinn weiter krabbelt sei ebenfalls q. Insgesamt muss also $p + 2q = 1$ sein. Ein Zeitschritt in diesem System wird also durch folgende stochastische Matrix beschrieben:

$$P \quad = \quad \begin{pmatrix} p & q & 0 & q \\ q & p & q & 0 \\ 0 & q & p & q \\ q & 0 & q & p \end{pmatrix}$$

Der Anfangszustand ist $(1,0,0,0)$. Als erstes fragen wir uns mit welcher Wahrscheinlichkeit die Spinne nach 2 Schritten immer noch oder wieder in Ecke 1 angekommen ist. Dies lässt sich offenbar beantworten, durch

$$P^2 \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} p^2 + 2q^2 \\ 2pq \\ 2q^2 \\ 2pq \end{pmatrix}$$

Die Antwort ist also $p^2 + 2q^2$ - dies stimmt mit unserer Intuition überein, die Spinne kann entweder zwei mal sitzen geblieben sein p^2 oder hin und wieder zurück gekrabbelt sein $2q^2$. Der Faktor 2 kommt daher, da es zwei Möglichkeiten für den Drehsinn gibt. Analog lassen sich auf diese Weise die Aufenthaltswahrscheinlichkeiten nach beliebig vielen Schritten für jede Ecke berechnen. Die Berechnung wird aber für große Schrittzahlen k leichter, wenn man P zuvor diagonalisiert (mit $\lambda := p - 2q$ gilt):

$$\begin{pmatrix} p & q & 0 & q \\ q & p & q & 0 \\ 0 & q & p & q \\ q & 0 & q & p \end{pmatrix}^k = \begin{pmatrix} 0 & -1 & -1 & 1 \\ -1 & 0 & 1 & 1 \\ 0 & 1 & -1 & 1 \\ 1 & 0 & 1 & 1 \end{pmatrix} \begin{pmatrix} p^k & 0 & 0 & 0 \\ 0 & p^k & 0 & 0 \\ 0 & 0 & \lambda^k & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 0 & -1 & -1 & 1 \\ -1 & 0 & 1 & 1 \\ 0 & 1 & -1 & 1 \\ 1 & 0 & 1 & 1 \end{pmatrix}^{-1}$$

Als nächstes interessieren wir uns für die Frage in welcher Ecke die Spinne nach langer Zeit ($k \to \infty$) am liebsten sitzt. Wie wir wissen ist die Grenzverteilung ℓ gerade der Eigenvektor zum Eigenwert 1, wir müssen also die Gleichung $P\ell = \ell$ lösen. Und für diese ergibt sich

$$P\ell = \ell \in \Delta_4 \quad \Longrightarrow \quad \ell = \frac{1}{4} \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}$$

D.h. alle Ecken sind gleich wahrscheinlich. Wie sollte es auch anders sein? Nach unendlich vielen Schritten spielt der Startpunkt keine Rolle mehr (Ergodensatz) und dann sind alle Ecken gleichberechtigt.

Beispiel 9.10:
Für ihr Risikomanagement verwenden Banken (bis zu) 22 verschiedene Bonitätsstufen bei ihren Kreditnehmern - von AAA bis D (wobei D für Kreditausfall steht). Über die Jahre hinweg wurden die Übergangswahrscheinlichkeiten von einer Kreditstufe in eine andere Kreditstufe gesammelt. Will ein Kunde der Bonität s also einen Kredit über m Jahre Laufzeit, dann berechnet die Bank die Ausfallwahrscheinlichkeit des Kredits mit Hilfe von $P^m e_s$. Und anhand der Ausfallwahrscheinlichkeit werden die Zinsen so berechnet, dass der Erwartungswert immer noch für die Bank positiv ist.

Um ein konkretes Beispiel geben zu können, betrachten wir 4 Bonitätsstufen A bis D. Die (willkürlich gewählte) Übergangsmatrix sei:

$$P = \begin{pmatrix} 0.80 & 0.05 & 0.10 & 0.00 \\ 0.10 & 0.75 & 0.10 & 0.00 \\ 0.05 & 0.15 & 0.60 & 0.10 \\ 0.05 & 0.05 & 0.20 & 0.90 \end{pmatrix}$$

Man beachte, dass in diesem Fall also eine 10% Chance besteht, dass ein Kredit der Stufe $D = 4$ wieder bedient werden kann (zu $C = 3$ wird). Diese Matrix besitzt die Eigenwerte $\lambda_1 = 1$, $\lambda_2 = 0.85$, $\lambda_3 = 0.7$ und $\lambda_4 = 0.5$ und die zugehörigen Eigenvektoren (in den Spalten von T)

$$T = \begin{pmatrix} 2 & 1 & 2 & 2 \\ 2 & 1 & -2 & 2 \\ 3 & 0 & -1 & -7 \\ 8 & -2 & 1 & 3 \end{pmatrix}$$

Sagen wir ein Kunde mit Bonität $B = 2$ kommt zur Bank und möchte einen Kredit von K Euro bei m Jahren Laufzeit haben. Er befindet sich also zu Beginn im Zustand $b = e_2 = (0, 1, 0, 0)$. Der Kredit fällt im k-ten Schritt aus (wobei $k \in 1 \ldots m$), wenn er am Ende des Schrittes auf Stufe $D = 4$ gelandet ist. Die Wahrscheinlichkeit dafür ist $a(k) := (P^k b)[4]$. Beispielsweise ist hier $a(1) = 5\%$ (eine unrealistisch hohe Zahl). Im zweiten Schritt ist diese aber schon auf $a(2) = 11.5\%$ gestiegen. In diesem Modell konvergiert $a(k)$ für $k \to \infty$ gegen $8/15 \approx 53.3\%$, wie wir an dem Eigenvektor zu 1 ablesen können.

Bleiben wir unrealistisch: die Bank gewährt den Kredit (wahrscheinlich hatte der Kunde einen schicken Anzug). Dann zahlt er am Ende jeden Schrittes eine Rate R des Kredites ab. Wie hoch muss die Bank die Rate ansetzen? Nehmen wir an, ein Kredit der $D = 4$ erreicht fällt aus (ohne Wiederkehr) und vernachlässigen wir die Inflation. Kann der Kredit dann k Schritte lang bedient werden beträgt der Gewinn der Bank $G(k) = kR - K$. Der Erwartungswert des Gewinns wäre bei maximal m Schritten Laufzeit also

$$E(G) = \sum_{k=0}^{m} p(k)G(k) = R\sum_{k=1}^{m} kp(k) - K$$

wobei $p(k)$ die Wahrscheinlichkeit dafür ist, dass der Kredit im $(k+1)$-ten Schritt ausfällt, aber den k-ten Schritt erreicht hat. Es ist also $p(0) = a(1)$ und $p(1) = (1 - a(1)) \cdot a(2)$ und $p(2) = (1 - a(1))(1 - a(2)) \cdot a(3)$ so weiter bis $p(m-1) = (1 - a(1)) \ldots (1 - a(m-1)) \cdot a(m)$ und schließlich noch $p(m) = (1 - a(1)) \ldots (1 - a(m))$. Allgemein gilt für $k \in 1 \ldots (m-1)$

$$p(k) = \frac{p(k-1)(1 - a(k))a(k+1)}{a(k)}$$

Da die Ausfallwahrscheinlichkeiten $a(k) = (P^k b)[D]$ berechnet werden können, kann die Bank auch rekursive $p(k)$ berechnen. Damit die Bank keinen Verlust macht, muss gelten $E(G) \geq 0$ und dies lässt sich umformen zu einer Bedingung für die zu zahlenden Raten:

$$\frac{R}{K} \geq \frac{1}{\sum_{k=1}^{m} kp(k)}$$

Bemerkung 9.11:
Auch wenn es in der Literatur immer wieder zu sehen ist: die Hardy-Weinberg Gesetzte der Genetik sind *kein* Beispiel für Übergangsmatrizen! Es handelt sich dabei vielmehr um ein nicht-lineares (quadratisches) Problem. Wir betrachten 1 Vererbungsmerkmal mit zwei möglichen Ausprägungen (*Allelen*) - zum Beispiel den Augenfarben braun und blau (in diesem Fall ist braun dominant). Wir gehen davon aus, dass das Merkmal keinen Einfluss auf das Fortpflanzungsverhalten hat (keine *Selektion*), dass eine zweigeschlechtliche Fortpflanzung stattfindet und dass weibliche und männliche Nachkommen gleich häufig sind. Der Genotyp werde mit GG, Gg bzw. gg bezeichnet, wobei G für das dominante und g für das rezessive Allel steht. Dann gibt es 6 verschiedene Typen in der Population: (f,GG), (f,Gg), (f,gg), (m,GG), (m,Gg) und (m,gg) wobei f/m für das Geschlecht steht. Wir bezeichnen die Häufigkeiten dieser Typen in der Population mit f_1, f_2, f_3, m_1, m_2 und m_3 Aus jeder Paarung gehen nach den Mendel'schen Gesetzen Nachkommen einer bestimmten Häufigkeitsverteilung vor:

f	m	GG	Gg	gg
GG	GG	1	0	0
GG	Gg	1/2	1/2	0
GG	gg	0	1	0
Gg	GG	1/2	1/2	0
Gg	Gg	1/4	1/2	1/4
Gg	gg	0	1/2	1/2
gg	GG	0	1	0
gg	Gg	0	1/2	1/2
gg	gg	0	0	1

D.h. der reinerbige Genotyp GG wird durch die Paarung $f_1 m_1$ sicher und zu 50% durch $f_1 m_2$ und $f_2 m_1$ und zu 25% durch $f_2 m_2$ gezeugt. Man beachte, dass davon nur die Hälfte der Nachkommen weiblich (bzw. männlich) sind, aber auch 2 Nachkommen pro Paar geboren werden müssen, um die Population konstant zu halten. Die Verteilung $(f_1', f_2', f_3', m_1', m_2', m_3')$ in der nächsten Generation lautet daher:

$$\begin{aligned}
f_1' &= f_1 m_1 + \frac{1}{2}(f_1 m_2 + f_2 m_1) + \frac{1}{4} f_2 m_2 \\
f_2' &= f_1 m_3 + f_3 m_1 + \frac{1}{2}(f_1 m_2 + f_2 m_1 + f_2 m_2 + f_2 m_3 + f_3 m_2) \\
f_3' &= f_3 m_3 + \frac{1}{2}(f_2 m_3 + f_3 m_2) + \frac{1}{4} f_2 m_2
\end{aligned}$$

Desgleichen gilt natürlich $m_i' = f_i'$ da beide Geschlechter gleich häufig geboren werden. Damit dies wieder eine Verteilung wird muss man allerdings noch normieren, d.h. multiplizieren mit $2/a$, wobei

$$a \;:=\; \sum_{i=1}^{3} f_i' \;=\; \sum_{i=1}^{3}\sum_{j=1}^{3} f_i m_j$$

Man sieht an diesem Bildungsgesetz, dass der Anfangszustand sehr wohl eine Rolle spielt - kommt nur gg in der Population von, dann wird niemals GG auftreten. Oder noch extremer: hat die Population keine Weibchen, so stirbt sie bereits in der nächsten Generation aus. Diese Markov-Prozeß ist also nicht *ergodisch*, er entwickelt sich nicht zu einem stabilen Gleichgewichtszustand und kann daher nicht von einer Übergangsmatrix stammen.

Dieser Prozess hat aber durchaus stabile Zustände $(f_1, f_2, f_, m_1, m_2, m_3)$. Offensichtlich sind $(1,0,0,1,0,0)/2$ und $(0,0,1,0,0,1)/2$ aber es gibt auch den nicht ganz so offensichtlichen stabilen Zustand $(1,2,1,1,2,1)/8$ in dem der gemischte Genotyp doppelt so häufig ist, wie die beiden reinerbigen.

Das Hardy-Weinberg-Gesetz besagt nun, dass dieses System nach spätestens zwei Schritten in einem stabilen Zustand angekommen ist. D.h. bezeichnet $f_i^{(k+1)} := (f_i^{(k)})'$ dann gilt für alle $k \geq 2$ bereits $f_i^{(k)} = f_i^{(2)}$. Dies lässt sich, mit ein bisschen Mühe, unmittelbar nachrechnen.

Kapitel 10

Mengenringe

Bei Laplace-Räumen konnte jedem Elementarereignis $s \in S$ eine Wahrscheinlichkeit $p(s) \in [0,1]$ zugeordnet werden. Beim Boule-Spiel aber rollt man eine Kugel und fragt sich wo die Kugel zum Stehen kommt. Vereinfacht gesagt wäre der Ereignisraum also $S = \mathbb{R}^+$ ein Kontinuum. Wir haben also folgendes Problem: die Kugel kann an *überabzählbar* vielen Punkten landen und damit hat jeder einzelne Punkt $s \geq 0$ eine Wahrscheinlichkeit von $p(s) = 0$ einzutreten. Erst ein echter Abschnitt auf der Zahlengeraden $[a,b] \subseteq \mathbb{R}^+$ hat eine Wahrscheinlichkeit $p([a,b]) > 0$ einzutreten. Aus der *Schreibweise* $p(A)$ wird damit also eine *Notwendigkeit*.

Wir müssen also dazu übergehen nicht mehr den Elementarereignissen $s \in S$ selbst eine Wahrscheinlichkeit zuzuordnen, sondern einer ganzen Teilmenge $A \subseteq S$. Und die Kolmogorov-Eigenschaften werden wir als Forderungen an p stellen müssen. Z.B. sollte p als Wahrscheinlichkeitsfunktion additiv sein $p(A \cup B) = p(A) + p(B)$ (für $A \cap B = \emptyset$). Damit ist eine Wahrscheinlichkeitsfunktion $p : A \mapsto p(A)$ also ein Spezialfall einer Volumenfunktion $\mu : A \mapsto \mu(A)$, die den Teilmengen eines Raumes S ihr Volumen zuordnen soll. So gesehen ist die Wahrscheinlichkeitstheorie also in die Maßtheorie übergegangen und erbt erst einmal die 2 Schwierigkeiten der Maßtheorie:

1. Problem: Es ist klar, dass es Mengen gibt, die unendlich großes Volumen haben, z.B. hat die ganze Ebene $S = \mathbb{R}^2$ unendlich großen Flächeninhalt. Die allgemeine Maßtheorie muss also den Fall $\mu(A) = \infty$ erlauben. In der Stochastik kommt das aber nicht vor, da wir maximal $p(S) = 1$ bekommen.

2. Problem: Damit nicht genug: es wäre halb so wild, wenn man zumindest jeder Teilmenge $A \subseteq \mathbb{R}^n$ immer eine Wahrscheinlichkeit $p(A)$ zuordnen könnte. Doch da kommt auch schon das nächste Problem daher: es ist möglich eine Kugel so geschickt zu zerlegen, dass man aus den Bruchstücken zwei gleiche Kugeln zusammensetzen kann! Ist K die Kugel, dann wäre also $\mu(K) = 2\mu(K)$ und damit $\mu(K) = 0$, was keinen Sinn ergibt. Das liegt daran, dass die Bruchstücke so fein zersplittert sind, dass man ihnen kein sinnvolles Volumen mehr zuordnen kann. Wir müssen uns also mit der Tatsache abfinden, dass nicht jeder Teilmenge $A \subseteq \mathbb{R}^3$ ein sinnvolles Volumen zugeordnet werden kann. Und damit müssen zunächst diejenigen Mengen aussortieren, die brav genug dafür sind.

Wir wollen hier die Maßtheorie noch nicht in voller Breite darstellen, sondern deren Konstruktionen nur im Endlichen nachvollziehen. So wie die Laplace-Räume ein erster Schritt in Richtung Maßtheorie waren, so sind die Mengenringe ein zweiter Schritt. Die Maßtheorie beschäftigt sich dann mit sogenannten σ-Algebren, die eine transfinite Version der Mengenringe sind (sie erlauben auch abzählbare Vereinigungen von Teilmengen). Darauf gehen wir aber erst im nächsten Kapitel ein.

Bemerkung 10.1: Banach-Tarski-Paradoxon:
Für die Interessierten hier ein paar Details zur Zerlegung der Kugel in Einzelteile, die durch Zusammensetzen zwei identische Kugeln ergeben: Zunächst betrachten wir die freie Gruppe mit 2 Erzeugern x und y. Das bedeutet die Elemente von $\mathcal{F}_2$ sind Ausdrücke der Form $a_1 a_2 \dots a_n$ wobei $a_i \in \{ x, x^{-1}, y, y^{-1} \}$ ist und x und x^{-1} bzw. y und y^{-1} nicht aufeinander folgen dürfen - die Worte sind gekürzt.

Formal erhält man $\mathcal{F}_2$ durch folgende Konstruktion: wir betrachten ein Alphabet bestehend aus 4 Buchstaben $\mathcal{A} := \{ u, u', v, v' \}$ und betrachten die Menge der Worte über diesem Alphabet:

$$\mathcal{A}^* \;:=\; \{ \varepsilon \} \cup \{ a_1 a_2 \dots a_n \mid 1 \leq n \in \mathbb{N}, a_i \in \mathcal{A} \}$$

(dabei ist ε das *leere Wort*). Offenbar ist $\mathcal{A}^*$ ein Monoid unter dem aneinanderhängen (Konkatenation) von Worten mit neutralem Element ε. Auf dieser Menge erklären wir die Relationen $uu' = \varepsilon$ und $vv' = \varepsilon$ (durch schneiden mit einer geeigneten Äquivalenzrelation). Dann ist $e = [\varepsilon]$, $x = [u]$ und $y = [v]$ mit den inversen Elementen $x^{-1} = [u']$ und $y^{-1} = [v']$. Die Verknüpfung wird durch die Konkatenation induziert: $[a][b] = [ab]$.

Man beachte, dass diese Gruppe eine bemerkenswerte Eigenschaften hat, die wir später ausnutzen werden: Bezeichnen wir die Menge der Worte mit dem Anfangsbuchstaben $a \in \{ x, x^{-1}, y, y^{-1} \}$ mit

$$F(a) \;:=\; \{ a_1 a_2 \dots a_n \in \mathbb{F}_2 \mid a_1 = a \}$$

dann darf der zweite Buchstabe a_2 dieser Worte nicht das Inverse von a sein, also: $a_2 \neq a^{-1}$ - sonst wäre das Wort ja nicht gekürzt. Ist $aa_2 \dots a_n \in F(a)$, dann beginnt $a^{-1} a a_2 \dots a_n = a_2 \dots a_n$ also mit dem Buchstaben $a_2 \neq a$. Daher gilt, z.B. $xF(x^{-1}) = \{ e \} \cup F(x^{-1}) \cup F(y) \cup F(y^{-1})$. Und damit erhalten wir

$$\mathcal{F}_2 \;=\; \{ e \} \cup F(x) \cup F(x^{-1}) \cup F(y) \cup F(y^{-1}) \;=\; F(x) \cup xF(x^{-1})$$

Als zweite Gruppe betrachten wir die Rotationen des $\mathbb{R}^3$, d.h. es sei $\mathcal{R} = \{ T \in \mathrm{mat}_3(\mathbb{R}) \mid T^*T = \mathbb{1} = TT^* \}$. Offenbar ist $\mathcal{R}$ eine Gruppe unter der Multiplikation von Matritzen. Wir betten nun $\mathcal{F}_2$ in $\mathcal{R}$ ein: Sei X die Rotation um die x_1-Achse um den Drehwinkel $\cos^{-1}(1/3)$ und Y die Rotation um die x_3-Achse um denselben Winkel. Wichtig ist hierbei, dass der Drehwinkel ein irrationales Vielfaches von π ist und die beiden Drehachsen senkrecht zueinander liegen. Dann erhalten wir einen injektiven Gruppenhomomorphismus durch Fortsetzung von

$$\mathcal{F}_2 \hookrightarrow \mathcal{R} \;:\; x \mapsto X \quad \text{und} \quad y \mapsto Y$$

D.h. wir haben ein isomorphes Bild von $\mathcal{F}_2$ in $\mathcal{R}$ gefunden. Und dieses erfüllt damit dieselbe, bemerkenswerte Zerlegung wie $\mathcal{F}_2$ selbst.

Sei nun $K := \{ k \in \mathbb{R}^3 \mid \|k\| \leq 1 \}$ die 3-dimensionale Kugel. Dann operiert $\mathcal{F}_2$ auf K vermöge der Rotationen X und Y. Nach dem Auswahlaxiom können wir aus jedem Orbit einen Punkt auswählen. Wir wählen also eine Teilmenge $J \subseteq K$ so dass

$$J \longleftrightarrow {}^{K}\!/_{\mathcal{F}_2} \ : \ j \mapsto \mathcal{F}_2 j$$

Damit definert man dann die folgenden Bruchstücke von K: $A_1 := F(x)J \cup J \cup B$, $A_2 := F(x^{-1})J \setminus B$, $A_3 := F(y)J$ und $A_4 := F(y^{-1})J$. Dabei ist

$$B \ := \ \bigcup_{i=1}^{\infty} x^{-i} J$$

Damit ist dann $K = A_1 \cup A_2 \cup A_3 \cup A_4 \cup N$ wobei $N \subseteq K$ noch eine kleine Nullmenge ist, die für das Volumen keine Rolle spielt. Für diese Mengen kann man dann zeigen, dass $xA_2 = A_2 \cup A_3 \cup A_4$ und $yA_4 = A_1 \cup A_2 \cup A_4$ ist und damit können wir zwei Kugeln aus diesen Bruchstücken zusammen setzen:

$$K \ = \ A_1 \cup xA_2 \cup N \quad \text{und} \quad K \ = \ A_3 \cup yA_4 \cup N$$

Definition 10.2:
Sei $S \neq \emptyset$ eine beliebige, nicht-leere Menge - wie immer interpretieren wir S als den Ereignisraum, in dem die Versuchsergebnisse liegen können. Und sei $\mathcal{A} \subseteq \mathcal{P}(S) = \{ A \mid A \subseteq S \}$ eine Menge von Teilmengen von S. Die Teilmengen $A \in \mathcal{A}$ von S interpretieren wir, als die zulässigen Ereignisbereiche. Dann definieren wir die folgenden Begriffe:

(i) Wir nennen $\mathcal{A}$ einen Mengenverband, falls $\emptyset$ in $\mathcal{A}$ enthalten ist und $\mathcal{A}$ abgeschlossen unter Vereinigung und Differenz von Mengen ist:

(1) $\emptyset \in \mathcal{A}$
(2) $A, B \in \mathcal{A} \implies B \setminus A \in \mathcal{A}$
(3) $A, B \in \mathcal{A} \implies A \cup B \in \mathcal{A}$

Und setzen wir $C := A \cup B \in \mathcal{A}$, dann gilt auch $A \cap B = C \setminus ((C \setminus A) \cup (C \setminus B)) \in \mathcal{A}$, so dass in diesem Fall weiterhin gilt:

(4) $A, B \in \mathcal{A} \implies A \cap B \in \mathcal{A}$

(ii) Wir nennen $\mathcal{A}$ einen Mengenring, falls $\emptyset$ in $\mathcal{A}$ enthalten ist und $\mathcal{A}$ abgeschlossen unter Vereinigung und Komplement ($\complement A := S \setminus A$) von Mengen ist. D.h. falls gilt:

(1) $\emptyset \in \mathcal{A}$
(2) $A \in \mathcal{A} \implies \complement A \in \mathcal{A}$
(3) $A, B \in \mathcal{A} \implies A \cup B \in \mathcal{A}$

Wegen $S = \mathsf{C}\emptyset \in \mathcal{A}$ und $A \cap B = \mathsf{C}(\mathsf{C}A \cup \mathsf{C}B) \in \mathcal{A}$ bzw. $B \setminus A = B \cap \mathsf{C}A \in \mathcal{A}$ erfüllt ein Mengenring stets die weiteren Eigenschaften:

(4) $S \in \mathcal{A}$

(5) $A, B \in \mathcal{A} \implies A \cap B \in \mathcal{A}$

(6) $A, B \in \mathcal{A} \implies B \setminus A \in \mathcal{A}$

(iii) Ist $\mathcal{A}$ ein Mengenverband, dann nennen wir eine Funktion $\mu : \mathcal{A} \to \mathbb{R}^+$ einen **Inhalt** auf $\mathcal{A}$, falls gilt:

(1) $\mu(\emptyset) = 0$

(2) $A, B \in \mathcal{A}$ mit $A \cap B = \emptyset \implies \mu(A \cup B) = \mu(A) + \mu(B)$

In diesem Fall nennen wir das Tripel $(S, \mathcal{A}, \mu)$ auch einen **Raum mit Inhalt**. Man beachte, dass die Menge $\mathbb{R}^+$ in der der Inhalt gemessen wird durch eine beliebige kommutative Gruppe ersetzt werden kann.

(iv) Ist $\mathcal{A}$ ein Mengenring, dann nennen wir eine Funktion $p : \mathcal{A} \to \mathbb{R}^+$ eine **Wahrscheinlichkeitsfunktion** auf $\mathcal{A}$, falls gilt:

(1) $p(S) = 1$

(2) $A, B \in \mathcal{A}$ mit $A \cap B = \emptyset \implies p(A \cup B) = p(A) + p(B)$

In diesem Fall nennen wir das Tripel $(S, \mathcal{A}, p)$ auch einen **Raum mit Wahrscheinlichkeit**. Es macht keinen Sinn $\mathbb{R}^+$ zu verallgemeinern. Wir nennen einen Punkt $s \in S$ ein *Elementarereignis* und interpretieren diesen als mögliches Ergebnis eines Zufallsexperiments. Eine Teilmenge $A \in \mathcal{A}$ heißt *Ereignis* und wir interpretieren $p(A)$ als *Wahrscheinlichkeit*, dass das Ergebnis des Zufallsexperiments in A liegt.

Bemerkung 10.3:

(i) Ein Mengenring ist also insbesondere ein Mengenverband. Ist umgekehrt $\mathcal{A}$ ein Mengenverband mit $S \in \mathcal{A}$, dann ist $\mathcal{A}$ auch ein Mengenring. D.h. es gilt die Äquivalenz:

$$\mathcal{A} \text{ Mengenring} \iff \mathcal{A} \text{ Mengenverband und } S \in \mathcal{A}$$

(ii) Ist $\mu : \mathcal{A} \to \mathbb{R}^+$ ein Inhalt auf dem Mengenverband $\mathcal{A}$, dann hat μ weiterhin die folgenden Eigenschaften:

(1) $\mu(\emptyset) = 0$ folgt eigentlich schon aus (2), da $\emptyset \cap \emptyset = \emptyset = \emptyset \cup \emptyset$ so dass aufgrund von (2) gilt $\mu(\emptyset) = \mu(\emptyset \cup \emptyset) = \mu(\emptyset) + \mu(\emptyset)$

(3) $A, B \in \mathcal{A} \implies \mu(B \setminus A) = \mu(B) - \mu(A \cap B)$, denn B besitzt eine disjunkte Zerlegung in $B = (A \cap B) \cup (B \setminus A)$ so dass aufgrund von (2) gilt: $\mu(B) = \mu(A \cap B) + \mu(B \setminus A)$

(4) $A, B \in \mathcal{A} \implies \mu(A \cup B) = \mu(A) + \mu(B) - \mu(A \cap B)$, denn $A \cup B$ besitzt die disjunkte Zerlegung $A \cup B = (A \setminus B) \cup (A \cap B) \cup (B \setminus A)$. Und daher gilt $\mu(A \cup B) = \mu(A \setminus B) + \mu(A \cap B) + \mu(B \setminus A)$. Mit (3) also: $\mu(A \cup B) = \mu(A) - \mu(A \cap B) + \mu(A \cap B) + \mu(B) - \mu(B \cap A) = \mu(A) + \mu(B) - \mu(A \cap B)$

(5) $A, B \in \mathcal{A}$ mit $A \subseteq B \implies \mu(A) \leq \mu(B)$, denn $A \subseteq B$ bedeutet $A \cap B = A$ und damit $0 \leq \mu(B \setminus A) = \mu(B) - \mu(A \cap B) = \mu(B) - \mu(A)$

(iii) Ist $p : \mathcal{A} \to \mathbb{R}^+$ sogar eine Wahrscheinlichkeitsfunktion auf dem Mengenring $\mathcal{A}$, so ist p trivialer Weise ein Inhalt auf $\mathcal{A}$ und erfüllt damit ebenfalls die obigen Eigenschaften. Da wir S aber disjunkt zerlegen können, in $S = A \cup \complement A$ folgt aus (2) bereits $1 = p(S) = p(A) + p(\complement A)$ und damit für alle Teilmengen $A \in \mathcal{A}$ auch

$$p(\complement A) \;=\; 1 - p(A)$$

(iv) Man führt für Räume mit Wahrscheinlichkeit $(S, \mathcal{A}, p)$ genauso die Begriffe der *bedingten Wahrscheinlichkeit* und *Unabhängigkeit* ein, wie in (7.5) für Laplace-Räume. Für diese gelten dann wortgleich wieder der Satz der totalen Wahrscheinlichkeit (7.6) die Pfadregel (7.8) und der Satz von Bayes (7.9). Einzig von allen beteiligten Mengen S_i und $A \subseteq S$ müssen wir natürlich verlangen, dass sie aus $\mathcal{A}$ stammen.

(v) Der Name *Mengenring* kommt nicht von ungefär: tatsächlich wird $\mathcal{A}$ zu einem kommutativen Ring unter der *symmetrischen Differenz* Δ als Addition und dem Schnitt $\cap$ als Multiplikation:

$$\begin{aligned} A + B &:= A \,\Delta\, B = (A \setminus B) \cup (B \setminus A) \\ A \cdot B &:= A \cap B \end{aligned}$$

Das Nullelement ist dann $0 = \emptyset$ und das Einselement $1 = S$. Dieser Ring hat Charakteristik 2 und für alle $A \in \mathcal{A}$ ist $A + 1 = \complement A$ und $A^2 = A$. Die Ideale $\mathcal{I} \subseteq \mathcal{A}$ sind dann genau die Teilmengen von $\mathcal{A}$, für die gilt: (1) $\emptyset \in \mathcal{I}$, (2) $I, J \in \mathcal{I} \implies I \cup J \in \mathcal{I}$ und (3) für alle $A \in \mathcal{A}$, $I \in \mathcal{I}$ gilt $A \subseteq I \implies A \in \mathcal{I}$.

(vi) Ein Raum mit Inhalt $(S, \mathcal{A}, \mu)$ ist die ideale Grundlage dafür mit Integration zu beginnen: Eine Funktion $e : S \to \mathbb{R}$ heißt **einfach**, falls sie einen endlichen Wertebereich hat und ihre Fasern Mengen aus $\mathcal{A}$ sind, d.h. falls gilt:

(1) $e(S) = \{\, e(s) \mid s \in S \,\} \subseteq \mathbb{R}$ ist endlich, und

(2) $0 \neq x \in e(S) \implies e^{-1}(x) = \{\, s \in S \mid e(s) = x \,\} \in \mathcal{A}$

Die Menge der einfachen Funktionen $\mathcal{E}(S, \mathbb{R}) = \{\, e : S \to \mathbb{R} \mid e \text{ einfach} \,\}$ ist dann eine $\mathbb{R}$-Unter-halbalgebra von $\mathcal{F}(S, \mathbb{R})$. D.h. $\mathcal{E}(S, \mathbb{R})$ ist ein Unter-vektorraum von $\mathcal{F}(S, \mathbb{R})$, der sogar abgeschlossen unter (punktweiser) Multiplikation ist. Dies gilt sogar, wenn wir $\mathbb{R}$ durch einen beliebigen kommutativen Ring R ersetzen. Und das **Integral** einer einfachen Funktion e wird dann definiert, durch:

$$\int e \, d\mu \;:=\; \sum_{x \in e(S)} x \cdot \mu(e^{-1}(x))$$

Die Integrationstheorie beginnt dann weitere Funktionen $f : S \to \mathbb{R}$ so durch einfache Funktionen $e_n : S \to \mathbb{R}$ zu approximieren, dass die Folge der Integrale der e_n wieder (unabhängig von der Wahl der e_n) in $\mathbb{R}$ konvergiert. Der Grenzwert dieser Integrale wird dann als Integral von f bezeichnet:

$$\int f\, d\mu \;:=\; \lim_{n\to\infty} \int e_n d\mu \;\text{ wobei } (e_n) \to f$$

Satz 10.4:

(i) Sei S eine beliebige Menge und $\mathcal{U} \subseteq \mathcal{P}(S)$ irgend eine Sammlung von Teilmengen von S. Dann bezeichnen wir:

$$[\mathcal{U}] \;:=\; \left\{ \bigcup_{i=1}^{m} \bigcap_{i=1}^{n(i)} U_{i,j} \;\middle|\; 1 \le m, n(i) \in \mathbb{N},\; U_{i,j} \in \mathcal{U} \right\}$$

Dann ist $\mathcal{U} \subseteq [\mathcal{U}] \subseteq \mathcal{P}(S)$ und dazu ist $[\mathcal{U}]$ sowohl $\cup$-stabil als auch $\cap$-stabil. D.h. für alle $V, W \in [\mathcal{U}]$ gilt $V \cup W \in [\mathcal{U}]$ und $V \cap W \in [\mathcal{U}]$.

(ii) Sei $S \neq \emptyset$ eine nicht-leere Menge und zu jedem $i \in I \neq \emptyset$ sei ein Mengenverband $\mathcal{A}_i \subseteq \mathcal{P}(S)$ über S gegeben. Dann ist der Schnitt über alle $\mathcal{A}_i$ wieder ein Mengenverband über S:

$$\bigcap_{i\in I} \mathcal{A}_i \;\subseteq\; \mathcal{P}(S) \quad \text{ist ein Mengenverband}$$

Ist $\mathcal{U} \subseteq \mathcal{P}(S)$ eine beliebige Familie von Teilmengen von S, dann können wir insbesondere den von $\mathcal{U}$ *erzeugten* Mengenverband (als den kleinsten Mengenverband auf S, der $\mathcal{U}$ enthält) definieren:

$$\langle \mathcal{U} \rangle_{\mathrm{MV}} \;:=\; \bigcap \{\, \mathcal{A} \subseteq \mathcal{P}(S) \mid \mathcal{A} \text{ ist ein Mengenverband, mit } \mathcal{U} \subseteq \mathcal{A} \,\}$$

(iii) Ist analog zu (ii) zu jedem $i \in I \neq \emptyset$ ein Mengenring $\mathcal{A}_i \subseteq \mathcal{P}(S)$ über S gegeben, dann ist der Schnitt über alle $\mathcal{A}_i$ wieder ein Mengenring:

$$\bigcap_{i\in I} \mathcal{A}_i \;\subseteq\; \mathcal{P}(S) \quad \text{ist ein Mengenring}$$

Und zu $\mathcal{U} \subseteq \mathcal{P}(S)$ definieren wir wieder den von $\mathcal{U}$ *erzeugten* Mengenring (als den kleinsten Mengenring auf S, der $\mathcal{U}$ enthält):

$$\langle \mathcal{U} \rangle_{\mathrm{MR}} \;:=\; \bigcap \{\, \mathcal{A} \subseteq \mathcal{P}(S) \mid \mathcal{A} \text{ ist ein Mengenring, mit } \mathcal{U} \subseteq \mathcal{A} \,\}$$

(iv) Sei wieder $\mathcal{U} \subseteq \mathcal{P}(S)$, dann bezeichnen wir $\mathcal{Q} := [\mathcal{U}] \cup \{\, \emptyset \,\}$ und $\mathcal{L} := \{\, Q \setminus P \mid P, Q \in \mathcal{Q} \,\}$. Dann können wir den von $\mathcal{U}$ erzeugten Mengenverband explizit angeben, als:

$$\langle \mathcal{U} \rangle_{\mathrm{MV}} \;=\; \left\{ \bigcup_{i=1}^{n} L_i \;\middle|\; 1 \le n \in \mathbb{N},\; L_i \in \mathcal{L} \right\}$$

Und bezeichnen wir $\mathcal{M} := \{\emptyset, S\} \cup \{U \mid U \in \mathcal{U}\} \cup \{\complement U \mid U \in \mathcal{U}\}$ dann können wir auch den von $\mathcal{U}$ erzeugten Mengenring explizit angeben:

$$\langle \mathcal{U} \rangle_{\mathrm{MR}} = [\mathcal{M}]$$

(v) Ist $(S, \mathcal{A}, \mu)$ ein Raum mit Inhalt und sind A_1 bis $A_n \in \mathcal{A}$ messbare Mengen, dann gilt für die Vereinigung dieser Mengen:

$$\mu\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{k=1}^{n} (-1)^{k+1} \sum_{I \subseteq 1\ldots n,\ \#I=k} \mu\left(\bigcap_{i \in I} A_i\right)$$

$$\mu\left(\bigcap_{i=1}^{n} A_i\right) = \sum_{k=1}^{n} (-1)^{k+1} \sum_{I \subseteq 1\ldots n,\ \#I=k} \mu\left(\bigcup_{i \in I} A_i\right)$$

Dabei ist die Summe über $\#I = k$ als Summe über alle Teilmengen I von $1 \ldots n$ mit k Elementen zu verstehen. D.h. die Summe läuft tatsächlich über die endliche Menge $\{I \subseteq 1 \ldots n \mid \#I = k\}$.

(vi) Sei nun $(S, \mathcal{A}, p)$ ein Raum mit Wahrscheinlichkeit und $B \in \mathcal{A}$ eine Teilmenge mit $p(B) \neq 0$, dann erhalten wir einen weiteren Raum mit Wahrscheinlichkeit $(B, \mathcal{A}_B, p_B)$ - der die *Restriktion* von $(S, \mathcal{A}, p)$ auf B genannt wird - durch

$$\begin{aligned} \mathcal{A}_B &:= \{A \cap B \mid A \in \mathcal{A}\} \\ p_B &: \mathcal{A}_B \to \mathbb{R}^+ : A \cap B \mapsto \frac{p(A \cap B)}{p(B)} \end{aligned}$$

(vii) Sei $(S, \mathcal{A}, \mu)$ ein Raum mit Inhalt, T eine beliebige Menge und schließlich $f : S \to T$ irgend eine Funktion. Dann definieren wir den *Push-forward* von $(S, \mathcal{A}, \mu)$ unter f, durch $(T, f_*\mathcal{A}, f_*\mu)$, wobei

$$\begin{aligned} f_*\mathcal{A} &:= \{B \subseteq T \mid f^{-1}(B) \in \mathcal{A}\} \\ f_*\mu &: f^*\mathcal{A} \to \mathbb{R}^+ : B \mapsto \mu(f^{-1}(B)) \end{aligned}$$

Und damit ist $(T, f_*\mathcal{A}, f_*\mu)$ dann wieder ein Raum mit Inhalt. Ist $(S, \mathcal{A}, p)$ sogar ein Raum mit Wahrscheinlichkeit, dann ist dessen Push-forward $(T, f_*\mathcal{A}, f_*p)$ wiederum ein Raum mit Wahrscheinlichkeit.

(viii) Ist nun $(T, \mathcal{B}, \nu)$ ein Raum mit Inhalt, S eine beliebige Menge und $f : S \twoheadrightarrow T$ ist eine surjektive Funktion, dann definieren wir den *Pull-back* von $(T, \mathcal{B}, \nu)$ unter f, durch $(S, f^*\mathcal{B}, f^*\nu)$, wobei

$$\begin{aligned} f^*\mathcal{B} &:= \{f^{-1}(B) \mid B \in \mathcal{B}\} \\ f^*\nu &: f^*\mathcal{B} \to \mathbb{R}^+ : A \mapsto \nu(f(A)) \end{aligned}$$

Dies ist wohldefiniert, da aus der Surjektivität von f folgt, dass für $A = f^{-1}(B)$ gilt: $f(A) = B \in \mathcal{B}$. Und damit ist $(S, f^*\mathcal{B}, f^*\nu)$ wieder ein Raum mit Inhalt. Und ist $(T, \mathcal{B}, q)$ sogar ein Raum mit Wahrscheinlichkeit, dann ist dessen Pull-back $(S, f^*\mathcal{B}, f^*q)$ wiederum ein Raum mit Wahrscheinlichkeit.

(ix) *Produktsatz*: Sind $(S, \mathcal{A}, \mu)$ und $(T, \mathcal{B}, \nu)$ zwei Räume mit Inhalt, dann bezeichnen wir die Menge $\mathcal{A}\times\mathcal{B} := \{\, A \times B \subseteq S \times T \mid A \in \mathcal{A},\ B \in \mathcal{B} \,\}$. Dann ist $\mathcal{A} \times \mathcal{B}$ $\cap$-stabil und der von $\mathcal{A} \times \mathcal{B}$ erzeugte Mengenverband ist gegeben, durch

$$\mathcal{A}\otimes\mathcal{B} \;:=\; \langle \mathcal{A}\times\mathcal{B} \rangle_{\mathrm{MV}} \;=\; \left\{ \bigcup_{i=1}^{n} A_i \times B_i \;\middle|\; 1 \le n \in \mathbb{N},\ A_i \in \mathcal{A},\ B_i \in \mathcal{B} \right\}$$

Auf diesem Mengenverband gibt es einen eindeutigen Inhalt $\mu \otimes \nu$ mit der Eigenschaft $(\mu\otimes\nu)(A\times B) = \mu(A)\cdot\nu(B)$ für alle $A \in \mathcal{A}$ und $B \in \mathcal{B}$. Wir nennen den Raum mit Inhalt $(S\times T, \mathcal{A}\otimes\mathcal{B}, \mu\otimes\nu)$ den *Produktraum* von $(S, \mathcal{A}, \mu)$ und $(T, \mathcal{B}, \nu)$. Sind beide $(S, \mathcal{A}, p)$ und $(T, \mathcal{B}, q)$ sogar Räume mit Wahrscheinlichkeit, dann ist der Produktraum wiederum ein Raum mit Wahrscheinlichkeit.

Beispiel 10.5:

(i) Der kleinste Mengenring überhaupt ist offenbar $\{\, \emptyset, S \,\}$. Auf diesem gibt es nur die Wahrscheinlichkeitsfunktion $p(\emptyset) = 0$ und $p(S) = 1$. Der größte Mengenring ist die Potenzmenge $\mathcal{P}(S)$ selbst.

(ii) Ist $A \subseteq S$ eine beliebige Teilmenge, dann ist der von $\{\, A \,\}$ erzeugte Mengenring gegeben, durch $\{\, \emptyset, A, \complement A, S \,\}$. Der erzeugte Mengenverband hingegen ist nur $\{\, \emptyset, A \,\}$.

(iii) Die bisherige Theorie der Laplace-Räume ist nur ein Spezialfall der Räume mit Wahrscheinlichkeit: Ist (S, p) ein Laplace-Raum, dann ist $(S, \mathcal{P}(S), p)$ ein Raum mit Wahrscheinlichkeit, wobei für eine beliebige Teilmenge $A \subseteq S$ $p(A)$ definiert wird, durch

$$p(A) \;:=\; \sum_{s\in A} p(s)$$

(iv) Ist $\mathcal{O}$ eine *Topologie* auf der Menge S (also die Menge aller offenen Mengen von S), dann ist $\mathcal{O}$ insbesondere $\cap$- und $\cup$-stabil (es sind sogar beliebige Vereinigungen erlaubt) und es sind $\emptyset$ und $S \in \mathcal{O}$. D.h. es ist $\mathcal{O} = [\mathcal{O}]$. Zumeist ist $\mathcal{O}$ aber kein Mengenring, da die Komplemente offener Mengen ja abgeschlossen sind. Der von $\mathcal{O}$ erzeugte Mengenring $\mathcal{B} := \langle \mathcal{O} \rangle_{\mathrm{MR}} = \langle \mathcal{O} \rangle_{\mathrm{MV}}$ wird als *Borel-Ring* auf S bezeichnet. Nach Satz (iv) ist $\mathcal{B}$ explizit gegeben, durch

$$\mathcal{B} \;=\; \left\{ \bigcup_{i=1}^{n} L_i \;\middle|\; 1 \le n \in \mathbb{N},\ L_i \in \mathcal{L} \right\}$$

Dabei wird $\mathcal{L} = \{\, V \setminus U \mid U, V \in \mathcal{O} \,\}$ als Menge der *lokal abgeschlossenen* Mengen bezeichnet. Eine Menge $B \in \mathcal{B}$ wird in der Topologie *konstruierbar* genannt.

(v) Das bekannteste Beispiel eines Inhalts ist sicherlich der von Lebesgue, der ausführlich in [Bauer, Maß- und Integrationstheorie, §4] vorgestellt wird. Wir geben hier nur eine Skizze der Konstruktion wieder: Zunächst bezeichnen wir für beliebige $a,\ b \in \mathbb{R}^d$ den *d-dimensionalen Quader* $[a,b[:= \left\{ x \in \mathbb{R}^d \mid \forall\, i \in 1 \ldots d : a_i \leq x_i < b_i \right\}$. Die Menge all dieser Quader bezeichnen wir mit

$$\mathcal{I}^d \;:=\; \left\{ [a,b[\mid a,b \in \mathbb{R}^d,\ \forall\, i \in 1 \ldots d : a_i \leq b_i \right\}$$

Offensichtlich ist $\mathcal{I}^d$ $\cap$-stabil und sind $I,\ J \in \mathcal{I}^d$, dann ist $J \setminus I$ eine endliche Vereinigung von Mengen aus $\mathcal{I}^d$. Damit erhalten wir einen Mengenverband durch

$$\mathcal{F}^d \;:=\; \left\{ \bigcup_{k=1}^{n} I_k \;\middle|\; 1 \leq n \in \mathbb{N},\ I_k \in \mathcal{I}^d \right\}$$

Auf diesem Mengenverband gibt es genau einen Inhalt $\lambda^d : \mathcal{F}^d \to \mathbb{R}^+$, für den die Quader folgendes Volumen erhalten:

$$\lambda^d([a,b[) \;=\; \prod_{i=1}^{d} (b_i - a_i)$$

Aus dieser Eigenschaft erkennt man aber (durch Induktion über d), dass λ^d nichts anderes ist, als das d-fache Produkt von λ^1, formal:

$$\lambda^d \;=\; \lambda^1 \otimes \ldots \otimes \lambda^1 \quad (d\text{-mal})$$

Der Fortsetzungssatz [Bauer, Maß- und Integrationstheorie, §5] erlaubt es dann diesen Inhalt λ^d auf eindeutige Weise zu einem Maß auf die von $\mathcal{F}^d$ erzeugte σ-Algebra fort zu setzten. Diese σ-Algebra ist aber nichts anderes, als die Borel-σ-Algebra von $\mathbb{R}^d$

$$\langle \mathcal{F}^d \rangle_\Sigma \;=\; \mathcal{B}(\mathbb{R}^d) \;:=\; \langle \mathcal{O}(\mathbb{R}^d) \rangle_\Sigma$$

(vi) Für die Spezialisten: auf jeder lokal-kompakten (insbesondere hausdorff'schen) Gruppe G gibt es ein (bis auf einen Normierungsfaktor) eindeutig bestimmtes, Maß τ, für dass für alle $g \in G$ und $A \in \mathcal{B}(G)$ gilt: $\tau(gA) = \tau(A)$. Man nennt τ das Haar'sche Maß von G. In diesem Sinne ist $\lambda^d = \tau$ genau das Haar'sche Maß der kommutativen Gruppe $\mathbb{R}^d$ (Gruppe unter der Addition).

Beweis:

(i) Offensichtlich ist $\mathcal{U} \subseteq [\mathcal{U}]$, da wir für beliebiges $U \in \mathcal{U}$ die Wahl $m = 1$, $n(1) = 1$ und $U_{1,1} = U$ treffen können. Die $\cap$-Stabilität ist direkt zu sehen: sei $U = \bigcup_i U_i$ wobei $U_i = \bigcap_j U_{i,j}$ mit $U_{i,j} \in \mathcal{U}$ und analog $V = \bigcup_k V_k$ wobei $V_k = \bigcap_\ell V_{k,\ell}$ mit $V_{k,\ell} \in \mathcal{U}$. Dann ist

$$U \cup V \;=\; \left(\bigcup_{i=1}^{m} U_i\right) \cup \left(\bigcup_{k=1}^{p} V_k\right) \;=\; \bigcup_{\alpha=1}^{m+p} \begin{cases} U_i & \text{für } \alpha \leq m \\ V_{k-m} & \text{für } \alpha > m \end{cases}$$

offensichtlich wieder in $[\mathcal{U}]$ enthalten. Auf dieselbe Art sieht man aber auch, dass $U_i \cap V_k$ ein Schnitt über Mengen aus $\mathcal{U}$ ist:

$$U_i \cap V_k = \left(\bigcap_{j=1}^{n(i)} U_{i,j}\right) \cap \left(\bigcap_{\ell=1}^{q(k)} V_{k,\ell}\right) = \bigcap_{\alpha=1}^{n(i)+q(k)} \begin{cases} U_{i,\alpha} & \text{für } \alpha \leq n(i) \\ V_{k,\alpha-n(i)} & \text{für } \alpha > n(i) \end{cases}$$

Insbesondere ist $U_i \cap V_k$ also wieder in $[\mathcal{U}]$ enthalten. Wir haben aber schon gezeigt, dass $[\mathcal{U}]$ $\cup$-stabil ist und daraus erhält man dann auch die $\cap$-Stabilität $U \cap V \in [\mathcal{U}]$, aufgrund von:

$$U \cap V \;=\; \left(\bigcup_{i=1}^{m} U_i\right) \cap \left(\bigcup_{k=1}^{p} V_k\right) \;=\; \bigcup_{i=1}^{m} \bigcup_{k=1}^{p} U_i \cap V_k$$

(ii) Da alle $\mathcal{A}_i$ Mengenverbände sind, ist stets $\emptyset \in \mathcal{A}_i$. Und da $\emptyset$ in allen $\mathcal{A}_i$ enthalten ist, gilt auch $\emptyset \in \bigcap_i \mathcal{A}_i$. Und sind A, $B \in \bigcap_i \mathcal{A}_i$ dann sind also A, $B \in \mathcal{A}_i$ für alle $i \in I$. Damit gilt dann aber auch stets $B \setminus A$ und $A \cup B \in \mathcal{A}_i$. Und da dies wieder für alle $i \in I$ gilt erhält man wiederum $B \setminus A$ und $A \cup B \in \bigcap_i \mathcal{A}_i$. Insgesamt ist $\bigcap_i \mathcal{A}_i$ also wieder ein Mengenverband.

Man beachte, dass $\mathcal{P}(S)$ trivialer Weise selbst ein Mengenverband (sogar Mengenring) über S ist. Ist also $\mathcal{U} \subseteq \mathcal{P}(S)$ beliebig, dann ist die Menge aller Mengenverbände $\mathcal{A}$ mit $\mathcal{U} \subseteq \mathcal{A}$ also nicht leer. Und damit ist $\langle \mathcal{U} \rangle_{\mathrm{MV}}$ wohldefiniert und wie oben gezeigt ein Mengenverband.

(iii) Sind alle $\mathcal{A}_i$ sogar Mengenringe, dann gilt stets $S \in \mathcal{A}_i$. Da dies für alle $i \in I$ gilt, folgt also $S \in \bigcap_i \mathcal{A}_i$. Nach (ii) ist $\bigcap_i \mathcal{A}_i$ also ein Mengenverband mit $S \in \bigcap_i \mathcal{A}_i$ und damit ein Mengenring.

(iv) Ist $\mathcal{U} \subseteq \mathcal{P}(S)$ beliebig, dann ist $[\mathcal{U}]$ nach (i) $\cup$- und $\cap$-stabil. Und damit ist auch offensichtlich $\mathcal{Q} = [\mathcal{U}] \cup \{\emptyset\}$ $\cup$- und $\cap$-stabil. Weiterhin gilt $\mathcal{U} \subseteq \mathcal{Q}$. Offenbar ist aber auch $\mathcal{Q} \subseteq \mathcal{L}$, da wir jedes $Q \in \mathcal{Q}$ schreiben können, als $Q = Q \setminus \emptyset \in \mathcal{L}$. Wir bezeichnen nun den von $\mathcal{U}$ erzeugten Mengenverband mit $\mathcal{A} := \langle \mathcal{U} \rangle_{\mathrm{MV}}$. Da $\mathcal{A}$ abgeschlossen unter $\cup$ und $\cap$ ist, haben wir $\mathcal{Q} \subseteq \mathcal{A}$. Und da $\mathcal{A}$ uch abgeschlossen unter der Differenz von Mengen ist ist auch $\mathcal{L} \subseteq \mathcal{A}$. Und damit bekommen wir dann wieder $[\mathcal{L}] \subseteq \mathcal{A}$, insgesamt also

$$\mathcal{U} \;\subseteq\; \mathcal{Q} \;\subseteq\; \mathcal{L} \;\subseteq\; [\mathcal{L}] \;\subseteq\; \mathcal{A}$$

Zu bemerken ist, dass $\mathcal{L}$ $\cap$-stabil ist, denn für $L = Q \setminus P$ und $L' = Q' \setminus P' \in \mathcal{L}$ gilt offenbar $L \cap L' = (Q \setminus P) \cap (Q' \setminus P') = (Q \cap Q') \setminus (P \cup P')$. Und da $\mathcal{Q}$ ja $\cap$- und $\cup$-stabil ist bedeutet das auch $L \cap L' \in \mathcal{L}$. Damit reduziert sich $[\mathcal{L}]$ aber auf die Menge

$$[\mathcal{L}] \;=\; \mathcal{W} \;:=\; \left\{ \bigcup_{i=1}^{n} L_i \;\middle|\; 1 \leq n \in \mathbb{N},\; L_i \in \mathcal{L} \right\}$$

Betrachten wir wieder $L = Q \setminus P$ und $L' = Q' \setminus P' \in \mathcal{L}$ dann gilt aber auch $L \setminus L' = (Q \setminus (P \cup Q')) \cup ((Q \cap P') \setminus P) \in \mathcal{W}$. Und damit ist $\mathcal{W}$ also auch abgeschlossen unter der Differenz von Mengen

$$\left(\bigcup_{i=1}^{m} L_i \right) \setminus \left(\bigcup_{j=1}^{n} L'_j \right) \;=\; \bigcup_{i=1}^{m} \bigcap_{j=1}^{n} L_i \setminus L'_j \;\in\; \mathcal{W}$$

Insgesamt ist $\mathcal{W}$ also $\cup$-stabil, abgeschlossen unter $\setminus$ und enthält $\emptyset$. Damit ist $\mathcal{W}$ ein Mengenverband, mit $\mathcal{U} \subseteq \mathcal{W}$, so dass $\mathcal{A} \subseteq \mathcal{W} = [\mathcal{L}]$. Die umgekehrte Inklusion haben wir aber bereits gezeigt, so dass insgesamt $\mathcal{A} = \mathcal{W}$ ist.

(iv) Sei nun $\mathcal{A} := \langle \mathcal{U} \rangle_{\mathrm{MR}}$ der von $\mathcal{U}$ erzeugte Mengenring. Nach Konstruktion ist $\mathcal{M}$ abgeschlossen unter Komplementen, d.h. für alle $M \in \mathcal{M}$ gilt auch $\complement M \in \mathcal{M}$ liegt. Außerdem ist klar, dass $\mathcal{U} \subseteq \mathcal{M} \subseteq \mathcal{A}$. Nach (i) ist $[\mathcal{M}]$ abgeschlossen unter Vereinigung und Schnitt und es gilt $S \in \mathcal{M} \subseteq [\mathcal{M}]$. Damit sieht man dann aber leicht, dass $[\mathcal{M}]$ abgeschlossen unter Komplementen ist: mit $M_{i,j} \in \mathcal{M}$ ist stets auch $\complement M_{i,j} \in \mathcal{M}$, so dass

$$\begin{aligned} \complement \left(\bigcup_{i=1}^{m} \bigcap_{j=1}^{n(i)} M_{i,j} \right) \;&=\; \bigcap_{i=1}^{m} \bigcup_{j=1}^{n(j)} \complement M_{i,j} \\ &=\; \bigcup_{j(1)=1}^{n(1)} \cdots \bigcup_{j(m)=1}^{n(m)} \bigcap_{i=1}^{m} \complement M_{i,j(i)} \end{aligned}$$

und da $[\mathcal{M}]$ abgeschlossen unter $\cap$ und $\cup$ ist, liegt diese Menge also wieder in $[\mathcal{M}]$. Damit ist $[\mathcal{M}]$ ein Mengenring mit $\mathcal{U} \subseteq [\mathcal{M}]$ und damit gilt $\mathcal{A} \subseteq [\mathcal{M}]$. Da aber auch $\mathcal{M} \subseteq \mathcal{A}$ gilt und $\mathcal{A}$ abgeschlossen unter $\cap$ und $\cup$ ist, erhalten wir damit auch $[\mathcal{M}] \subseteq \mathcal{A}$, zusammen also $\mathcal{A} = [\mathcal{M}]$, was zu beweisen war.

(vi) Zunächst ist klar, dass wegen $B \in \mathcal{A}$ auch $\mathcal{A}_B$ ein Mengenring ist: $B = B \cap B \in \mathcal{A}_B$ und für $A,\, A' \in \mathcal{A}$ ist auch $(A \cap B) \cup (A' \cap B) = (A \cup A') \cap B \in \mathcal{A}_B$. Schließlich ist das Komplement ja in B zu nehmen und deswegen gilt für $A \in \mathcal{A}$ wieder $\complement(A \cap B) = B \setminus (A \cap B) = B \setminus A = (B \setminus A) \cap B \in \mathcal{A}_B$. Es bleibt also zu zeigen, dass p_B wieder eine Wahrscheinlichkeitsfunktion ist: zunächst ist $B \cap B = B$ also $p_B(B \cap B) = p(B)/p(B) = 1$. Und sind $A_B = A \cap B$ und $A'_B = A' \cap B \in \mathcal{A}_B$ mit $A_B \cap A'_B = \emptyset$ dann gilt $p_B(A_B \cup A'_B) = p(A_B \cup A'_B)/p(B) = (p(A_B) + p(A'_B))/p(B) = p(A_B)/p(B) + p(A'_B)/p(B) = p_B(A_B) + p_B(A'_B)$. Insgesamt ist $(B, \mathcal{A}_B, p_B)$ ein Raum mit Wahrscheinlichkeit.

(vii) Zunächst müssen wir wieder zeigen, dass $\mathcal{B} := f_*\mathcal{A}$ wieder ein Mengenverband ist. Offenbar ist $f^{-1}(\emptyset) = \emptyset \in \mathcal{A}$ und damit auch $\emptyset \in \mathcal{B}$. Und sind B, $B' \in \mathcal{B}$ dann gilt wiederum $f^{-1}(B \cup B') = f^{-1}(B) \cup f^{-1}(B') \in \mathcal{A}$ so dass auch $B \cup B' \in \mathcal{B}$ ist. Schließlich gilt auch $B' \setminus B = B' \cap \complement B$ so dass auch $f^{-1}(B' \setminus B) = f^{-1}(B' \cap \complement B) = f^{-1}(B') \cap \complement f^{-1}(B) = f^{-1}(B') \setminus f^{-1}(B) \in \mathcal{A}$ ist, woraus $B' \setminus B \in \mathcal{B}$ folgt. Und genauso sieht man, dass $\nu := f_*\mu$ ein Inhalt auf $\mathcal{B}$ ist: sei $B \cap B' = \emptyset$, dann ist

$$\begin{aligned}
\nu(B \cup B') &= \mu(f^{-1}(B \cup B')) \;=\; \mu(f^{-1}(B) \cup f^{-1}(B')) \\
&= \mu(f^{-1}(B)) + \mu(f^{-1}(B')) - \mu(f^{-1}(B) \cap f^{-1}(B')) \\
&= \mu(f^{-1}(B)) + \mu(f^{-1}(B')) - \mu(f^{-1}(B \cap B')) \\
&= \mu(f^{-1}(B)) + \mu(f^{-1}(B')) - \mu(\emptyset) \\
&= \mu(f^{-1}(B)) + \mu(f^{-1}(B')) \;=\; \nu(B) + \nu(B')
\end{aligned}$$

Eigenschaft (1) von Inhalten folgt ohnehin aus Eigenschaft (2). Insgesamt ist $(T, \mathcal{B}, \nu)$ also ein Raum mit Inhalt. Und ist $(S, \mathcal{A}, \mu)$ sogar ein Raum mit Wahrscheinlichkeit, dann gilt $S \in \mathcal{A}$ so dass auch $f^{-1}(T) = S \in \mathcal{A}$ ist, was wiederum $T \in \mathcal{B}$ bedeutet. Und aufgrund von $\nu(T) = \mu(f^{-1}(T)) = \mu(S) = 1$ ist ν dann auch eine Wahrscheinlichkeitsfunktion.

(viii) Zunächst müssen wir wieder zeigen, dass $\mathcal{A} := f^*\mathcal{B}$ wieder ein Mengenverband ist. Wiederum ist $\emptyset = f^{-1}(\emptyset) \in \mathcal{A}$, da ja $\emptyset \in \mathcal{B}$ ist. Und betrachten wir $A = f^{-1}(B)$ und $A' = f^{-1}(B') \in \mathcal{A}$, dann findet man wie schon in (vi) wieder $A \cup A' = f^{-1}(B \cup B')$ und $A' \setminus A = f^{-1}(B' \setminus B)$, so dass $A \cup A'$ und $A' \setminus A \in \mathcal{A}$ wieder in $\mathcal{A}$ enthalten sind. Man beachte, dass hierfür die Surjektivität von f *nicht* benötigt wurde. Nun ist also zu zeigen, dass $\mu := f^*\nu$ ein Inhalt auf $\mathcal{A}$ ist: seien wieder $A = f^{-1}(B)$ und $A' = f^{-1}(B') \in \mathcal{A}$ mit $A \cap A' = \emptyset$. Dann ist auch $\emptyset = f(\emptyset) = f(A \cap A') = f(f^{-1}(B \cap B')) = B \cap B'$ (letzteres, da f surjektiv ist) und damit auch $\mu(A \cup A') = \nu(f(A \cup A')) = \nu(B \cup B') = \nu(B) + \nu(B') = \mu(A) + \mu(A')$. Eigenschaft (1) von Inhalten folgt ohnehin aus Eigenschaft (2). Insgesamt ist $(S, \mathcal{A}, \mu)$ also ein Raum mit Inhalt. Und ist $(T, \mathcal{B}, \nu)$ sogar ein Raum mit Wahrscheinlichkeit, dann gilt $S = f^{-1}(T) \in \mathcal{A}$ und aus der Surjektivität folgt auch $\mu(S) = \nu(f(S)) = \nu(T) = 1$, so dass μ dann auch eine Wahrscheinlichkeitsfunktion ist.

(ix) Der Beweis dieser Aussage ist sehr umfangreich, so dass wir hier nur eine Skizze geben wollen: zunächst einmal ist klar, dass für A, $A' \in \mathcal{A}$ und B, $B' \in \mathcal{B}$ gilt: $(A \times B) \cap (A' \times B') = (A \cap A') \times (B \cap B')$. Da $\mathcal{A}$ und $\mathcal{B}$ aber $\cap$-stabil sind, ist damit also $(A \times B) \cap (A' \times B') \in \mathcal{A} \times \mathcal{B}$. D.h. diese Menge ist ebenfalls $\cap$-stabil. Nun bezeichnen wir

$$\mathcal{P} \;:=\; \left\{ \bigcup_{i=1}^{n} A_i \times B_i \;\middle|\; 1 \le n \in \mathbb{N},\; A_i \in \mathcal{A},\; B_i \in \mathcal{B} \right\}$$

Man macht sich recht leicht klar (Skizze), dass $(A \times B) \cup (A' \times B')$ und $(A' \times B') \setminus (A \times B)$ endliche Vereinigungen von Mengen aus $\mathcal{A} \times \mathcal{B}$ sind, d.h. in $\mathcal{P}$ liegen. Daraus folgt dann weiter, dass $\mathcal{P}$ abgeschlossen unter $\cap$, $\cup$ und $\setminus$ ist, also ein Mengenverband ist. Mithin gilt $\langle \mathcal{A} \times \mathcal{B} \rangle_{\mathrm{MV}} = \mathcal{P}$.

Sind $\mathcal{A}$ und $\mathcal{B}$ Mengenringe, dann gilt natürlich auch $S \times T \in \mathcal{P}$ und damit ist $\mathcal{P}$ dann wieder ein Mengenring.

Ist $P = \bigcup_i A_i \times B_i \in \mathcal{P}$, dann hat P eine Zerlegung als endliche, disjunkte Vereinigung von Mengen aus $\mathcal{A} \times \mathcal{B}$. Man kann diese recht leicht gewinnen, indem man definiert:

$$\alpha : P \to \mathcal{P}(1 \dots n) : (s,t) \mapsto \{\, i \in 1 \dots n \mid s \in A_i \,\}$$

$$\omega : P \to \mathcal{P}(1 \dots n) : (s,t) \mapsto \{\, j \in 1 \dots n \mid t \in B_j \,\}$$

Man beachte, dass $\alpha(s,t)$ und $\omega(s,t)$ nie die leere Menge sein können, da $(s,t) \in P$ ja in mindestens einem der $A_i \times B_i$ enthalten sein muss. Deren Fasern sind nun von der Gestalt (für $\emptyset \neq I,\, J \subseteq 1 \dots n$)

$$A_I \;:=\; \alpha^{-1}(I) \;=\; \left(\bigcap_{i \in I} A_i\right) \setminus \left(\bigcup_{i \notin I} A_i\right) \;\in\; \mathcal{A}$$

$$B_J \;:=\; \omega^{-1}(J) \;=\; \left(\bigcap_{j \in J} B_j\right) \setminus \left(\bigcup_{j \notin J} B_j\right) \;\in\; \mathcal{B}$$

Als Fasern sind die A_I bzw. B_J paarweise disjunkt und wir haben damit schließlich eine disjunkte Zerlegung von P gefunden, als

$$P \;=\; \bigcup_{I \neq \emptyset} \bigcup_{J \neq \emptyset} A_I \times B_J$$

Damit kann man den Inhalt $\pi : \mathcal{P} \to \mathbb{R}^+$ definieren, indem man zu $P \in \mathcal{P}$ eine Zerlegung von $P = \bigcup_i \bigcup_j A_i \times B_j$ wählt in der sowohl die $A_i \in \mathcal{A}$, als auch die $B_j \in \mathcal{B}$ paarweise disjunkt sind. Für diese definiert man dann π durch:

$$\pi\left(\bigcup_{i=1}^{m} \bigcup_{j=1}^{n} A_i \times B_j\right) \;:=\; \sum_{i=1}^{m} \sum_{j=1}^{n} \mu(A_i) \cdot \nu(B_j)$$

Dabei ist π wohldefiniert: die Summe hängt nicht von der gewählten Zerlegung ab, denn: seien $P = \bigcup_i \bigcup_j A_i \times B_j = \bigcup_r \bigcup_s C_r \times D_s$ zwei solche Zerlegungen von P. Dann ist für alle $i \in 1 \dots m$ und $j \in 1 \dots n$ offenbar

$$\begin{aligned} A_i \times B_j \;&=\; (A_i \times B_j) \cap P \;=\; (A_i \times B_j) \cap \left(\bigcup_{r=1}^{u} \bigcup_{s=1}^{v} C_r \times D_s\right) \\ &=\; \bigcup_{r=1}^{u} \bigcup_{s=1}^{v} (A_i \cap C_r) \times (B_j \cap D_s) \end{aligned}$$

Durch Projektion auf die jeweiligen Koordinaten findet man daraus disjunkte Zerlegungen $A_i = \bigcup_r A_i \cap C_r$ bzw. $B_j = \bigcup_s B_j \cap D_s$. Und da μ und ν Inhalte sind bedeutet das

$$\mu(A_i) \;=\; \sum_{r=1}^{u} \mu(A_i \cap C_r) \;\text{ bzw. }\; \nu(B_j) \;=\; \sum_{s=1}^{v} \nu(B_j \cap D_s)$$

Insgesamt also $\mu(A_i) \cdot \nu(B_j) = (\sum_r \mu(A_i \cap C_r))(\sum_s \nu(B_j \cap D_r)) = \sum_r \sum_s \mu(A_i \cap C_r)\nu(B_j \cap D_r)$. Und damit liefert die erste Zerlegung von P den folgenden Inhalt unter π:

$$\pi\left(\bigcup_{i=1}^{m}\bigcup_{j=1}^{n} A_i \times B_j\right) \;=\; \sum_{i=1}^{m}\sum_{j=n}^{n}\sum_{r=1}^{u}\sum_{s=1}^{v} \mu(A_i \cap C_r)\nu(B_j \cap D_r)$$

Vertauscht man die Rollen der $A_i \times B_j$ und $C_r \times D_s$ so findet man in völliger Analogie zur gerade gegebenen Argumentation

$$\pi\left(\bigcup_{r=1}^{u}\bigcup_{s=v}^{n} C_r \times D_s\right) \;=\; \sum_{r=1}^{u}\sum_{s=1}^{v}\sum_{i=1}^{m}\sum_{j=n}^{n} \mu(A_i \cap C_r)\nu(B_j \cap D_r)$$

Da die Summenzeichen aber vertauschen, stimmen die beiden Ausdrücke überein und wir haben die Unabhängigkeit von der speziellen Zerlegung bewiesen.

Es bleibt nur noch die Eindeutigkeit von $\pi = \mu \otimes \nu$ zu zeigen. Dies liegt daran, dass π auf der $\cap$-stabilen Menge $\mathcal{A} \times \mathcal{B}$ festgelegt wurde. Der Beweis der Aussage verwendet aber Eigenschaften von Dynkin-Systemen und ist überraschend trickreich. Wir verweisen wieder auf [Zeidler, Algebraic Statistics, auf www.wuala.com/zeidlerweb] für einen Beweis.

(v) Wir beweisen die Formel durch Induktion über die Anzahl n der Teilmengen. Der Fall $n = 1$ ist trivial und $n = 2$ wurde in einer Bemerkung bereits bewiesen, wir betrachten nun also $n \geq 2$ und $A_1, A_2, \ldots, A_{n+1} \in \mathcal{A}$, wobei wir die Bezeichnung $B := A_2 \cup \cdots \cup A_{n+1}$ einführen. Wie gesehen, gilt ja: $\mu(A_1 \cup A_2 \cup \cdots \cup A_{n+1}) = \mu(A_1 \cup B) = \mu(A_1) + \mu(B) - \mu(A_1 \cap B)$. Wir wenden nun die Induktionsvoraussetzung auf $\mu(B)$ und $\mu(A_1 \cap B)$ an, womit sich jeweils ergibt:

$$\begin{aligned}
\mu(B) = \;& \sum_{k=1}^{n}(-1)^{k+1} \sum_{I \subseteq 2\ldots(n+1),\ \#I=k} \mu\left(\bigcap_{i\in I} A_i\right) \\
= \;& \sum_{k=1}^{n}(-1)^{k+1} \sum_{\substack{I \subseteq 1\ldots(n+1),\\ 1\notin I,\ \#I=k}} \mu\left(\bigcap_{i\in I} A_i\right) \\
= \;& \sum_{k=1}^{n+1}(-1)^{k+1} \sum_{\substack{I \subseteq 1\ldots(n+1),\\ 1\notin I,\ \#I=k}} \mu\left(\bigcap_{i\in I} A_i\right)
\end{aligned}$$

$$
\begin{aligned}
\mu(A_1 \cap B) = \; & \mu\left(\bigcup_{i=2}^{n+1} A_1 \cap A_i\right) \\
= \; & \sum_{k=1}^{n} (-1)^{k+1} \sum_{I \subseteq 2\ldots(n+1),\, \#I=k} \mu\left(\bigcap_{i\in I} A_1 \cap A_i\right) \\
= \; & \sum_{k=1}^{n} (-1)^{k+1} \sum_{\substack{I \subseteq 1\ldots(n+1),\\ 1\in I,\, \#I=k+1}} \mu\left(\bigcap_{i\in I} A_i\right) \\
& - \sum_{k=2}^{n+1} (-1)^{k+1} \sum_{\substack{I \subseteq 1\ldots(n+1),\\ 1\in I,\, \#I=k}} \mu\left(\bigcap_{i\in I} A_i\right) \\
= \; & \mu(A_1) - \sum_{k=1}^{n+1} (-1)^{k+1} \sum_{\substack{I \subseteq 1\ldots(n+1),\\ 1\in I,\, \#I=k}} \mu\left(\bigcap_{i\in I} A_i\right)
\end{aligned}
$$

Insgesamt finden wir also

$$
\begin{aligned}
\mu\left(\bigcup_{i=1}^{n} A_i\right) \; & = \; \mu(A_1) + \mu(B) - \mu(A_1 \cap B) \\
& = \; \mu(A_1) + \sum_{k=1}^{n+1} (-1)^{k+1} \sum_{\substack{I \subseteq 1\ldots(n+1),\\ 1\notin I,\, \#I=k}} \mu\left(\bigcap_{i\in I} A_i\right) \\
& \quad - \mu(A_1) + \sum_{k=1}^{n+1} (-1)^{k+1} \sum_{\substack{I \subseteq 1\ldots(n+1),\\ 1\in I,\, \#I=k}} \mu\left(\bigcap_{i\in I} A_i\right) \\
& = \; \sum_{k=1}^{n} (-1)^{k+1} \sum_{I \subseteq 1\ldots n,\, \#I=k} \mu\left(\bigcap_{i\in I} A_i\right)
\end{aligned}
$$

Die zweite Formel könnte genauso gut direkt durch Induktion über n bewiesen werden, aber es ist einfacher die erste Formel dafür zu verwenden: Da alle Teilmengen A_k in der Vereinigung $B := A_1 \cup \cdots \cup A_n$ enthalten sind, können wir die $A_k \in \mathcal{A}$ als Teilmengen von $A_k \subseteq B \in \mathcal{A}$ auffassen (das bedeutet $\complement A_k = B \setminus A_k$), dann können wir weiterhin berechnen:

$$
\mu\left(\bigcap_{k=1}^{n} A_k\right) \; = \; \mu\left(\complement \bigcup_{k=1}^{n} \complement A_k\right) \; = \; \mu(B) - \mu\left(\bigcup_{k=1}^{n} \complement A_k\right)
$$

Wenden wir die schon bewiesene erste Formel auf diesen Ausdruck an, dann erhalten wir daraus dann auch die zweite Formel:

$$
\begin{aligned}
& \mu\left(\bigcap_{k=1}^{n} A_k\right) \\
& = \mu(B) - \sum_{k=1}^{n} (-1)^{k+1} \sum_{I \subseteq 1\ldots n,\, \#I=k} \mu\left(\bigcap_{i\in I} \complement A_i\right)
\end{aligned}
$$

$$
\begin{aligned}
&= \mu(B) + \sum_{k=1}^{n} (-1)^k \sum_{I \subseteq 1\ldots n,\ \#I=k} \mu\left(\complement \bigcup_{i\in I} A_i\right) \\
&= \mu(B) + \sum_{k=1}^{n} (-1)^k \sum_{I \subseteq 1\ldots n,\ \#I=k} \left[\mu(B) - \mu\left(\bigcup_{i\in I} A_i\right)\right] \\
&= \mu(B) + \sum_{k=1}^{n} (-1)^k \binom{n}{k} \mu(B) - \sum_{k=1}^{n} (-1)^k \sum_{I \subseteq 1\ldots n,\ \#I=k} \mu\left(\bigcup_{i\in I} A_i\right) \\
&= \mu(B) \sum_{k=0}^{n} \binom{n}{k} (-1)^k 1^{n-k} + \sum_{k=1}^{n} (-1)^{k+1} \sum_{I \subseteq 1\ldots n,\ \#I=k} \mu\left(\bigcup_{i\in I} A_i\right) \\
&= \mu(B)\left((-1)+1\right)^n + \sum_{k=1}^{n} (-1)^{k+1} \sum_{I \subseteq 1\ldots n,\ \#I=k} \mu\left(\bigcup_{i\in I} A_i\right) \\
&= 0 + \sum_{k=1}^{n} (-1)^{k+1} \sum_{I \subseteq 1\ldots n,\ \#I=k} \mu\left(\bigcup_{i\in I} A_i\right)
\end{aligned}
$$

□

Die folgende Eigenschaft sieht zwar etwas bizarr aus, wird aber interessante Konsequenzen für die Unabhängigkeit von Ereignissen haben. Wir formulieren die Aussage in ihrer allgemeinen Form:

Proposition 10.6:
Sei $(S, \mathcal{A}, \mu)$ ein Raum mit Inhalt [eigentlich genügt es sogar, wenn $\mu : \mathcal{A} \to R$ ein Inhalt mit Werten in einem beliebigen Integritätsbereich R ist]. Und seien A, $A_1, \ldots, A_n \in \mathcal{A}$ messbare Mengen, so dass $A = A_1 \cup \cdots \cup A_n$ ist. Nehmen wir schließlich noch an, dass für alle $\emptyset \neq I \subseteq 1 \ldots n$ gilt:

$$\frac{\mu\left(\bigcap_{i\in I} A_i\right)}{\mu(A)} \;=\; \prod_{i\in I} \frac{\mu(A_i)}{\mu(A)}$$

Dann gibt es bereits ein $k \in 1 \ldots n$, so dass A_k denselben Inhalt hat, wie A

$$\exists\, k \in 1 \ldots n \;:\; \mu(A_k) \;=\; \mu(A)$$

Beweis:
Wir bezeichnen $a_k := \mu(A_k) \in R$ und $x := \mu(A) \in R$. Nach Voraussetzung und mit Hilfe von Satz (10.4.(v)) gilt aber

$$x \;=\; \mu\left(\bigcup_{j=1}^{n} A_j\right) \;=\; \sum_{r=1}^{n} \sum_{\#I=r} (-1)^{r-1} \mu\left(\bigcap_{i\in I} A_i\right)$$

Wir dividieren diese Gleichung durch $x = \mu(A)$ und ziehen es in die Summe

$$1 \;=\; \sum_{r=1}^{n} \sum_{\#I=r} (-1)^{r-1} \frac{\mu\left(\bigcap_{i\in I} A_i\right)}{\mu(A)}$$

Dann können wir die zweite Voraussetzung einsetzen. Um die entstehenden Ausdrücke leichter schreiben zu können sagen wir $I = \{ i_1, \ldots, i_r \}$, dann

$$\begin{aligned} 1 &= \sum_{r=1}^{n} \sum_{\#I=r} (-1)^{r-1} \prod_{i \in I} \frac{\mu(A_i)}{\mu(A)} \\ &= \sum_{r=1}^{n} \sum_{\#I=r} (-1)^{r-1} \frac{\prod_{i \in I} a_i}{x^r} \\ &= \sum_{r=1}^{n} \sum_{\#I=r} (-1)^{r-1} \frac{a_{i_1} \cdot \cdots \cdot a_{i_r}}{x^r} \end{aligned}$$

Wir multiplizieren diese Gleichung nun wieder mit x^n durch und bringen die lange Summe auf die linke Seite, dann erhalten wir weiter

$$x^n = \sum_{r=1}^{n} \sum_{\#I=r} (-1)^{r-1} a_{i_1} \cdot \cdots \cdot a_{i_r} x^{n-r}$$

$$x^n + \sum_{r=1}^{n} \sum_{\#I=r} (-1)^{r} a_{i_1} \cdot \cdots \cdot a_{i_r} x^{n-r} = 0$$

Dann erkennt man, dass es bei dem r-ten Summanden einfach um das r-te elementarsymmetrische Polynom in a_1 bis a_r handelt und damit lässt sich dieser Ausdruck mit dem Satz von Vieta umschreiben, zu

$$(x - a_1)(x - a_2) \cdot \cdots \cdot (x - a_n) = = 0$$

Da R als Integritätsbereich vorausgesetzt war, muss also einer der Faktoren $x - a_k$ verschwinden. D.h. es gibt ein $k \in 1 \ldots n$, so dass $a_k = x$ ist.

$\square$

Kapitel 11

Sigma-Algebren

In diesem Abschnitt präsentieren wir eine kurze Einführung in die klassische Maßtheorie, wie sie in zahlreichen Lehrwerken zu finden ist: Während Mengenringe nur endliche Mengenoperationen zulassen, erlauben wir jetzt auch abzählbare Vereinigungen (und damit auch Schnitte). Der Grund dafür liegt auf der Hand: wir wollen komplizierte Mengen unbekannten Volumens durch einfache Mengen bekannten Volumens ausschöpfen können. Und einen Kreis kann man eben nicht mit *endlich* vielen Quadern ausschöpfen.

Der Preis dafür ist aber, dass wir damit auch Mengen mit unendlichem Volumen zulassen müssen. Deswegen nehmen wir ∞ als möglichen Wert, den das Maß annehmen kann auf. Näheres dazu siehe (2.3). Des weiteren benötigen wir einen Konvergenzbegriff auf der Wertemenge des Maßes. Während es leicht möglich ist, Inhalte auf Mengenverbänden mit Werten in einem beliebigen kommutativen Ring zu betrachten, muss der Wertebereich eines Maßes auf einer σ-Algebra zumindest ein kommutativer Ring mit Metrik sein.

Definition 11.1:
Wie bei Mengenringen sei $S \neq \emptyset$ wieder eine beliebige, nicht-leere Menge und $\Sigma \subseteq \mathcal{P}(S)$ sei irgend eine Familie von Teilmengen von S. Dann definieren wir die folgenden Begriffe:

(i) Wir nennen Σ eine **σ-Algebra**, wenn S in Σ enthalten ist und Σ abgeschlossen unter Komplementen und abzähbaren Vereinigungen ist:

(1) $S \in \Sigma$
(2) $A \in \Sigma \implies \complement A \in \Sigma$
(3) $A : \mathbb{N} \to \Sigma \implies \bigcup_{n\in\mathbb{N}} A(n) \in \Sigma$

Wegen $\emptyset = \complement S \in \Sigma$ ist Σ damit insbesondere $\cup$-stabil, d.h. auch abgeschlossen unter endlichen Vereinigungen und damit ein Mengenring und damit auch Mengenverband. Durch Übergang zu Komplementen $\bigcap_n A(n) = \complement \bigcup_n \complement A(n)$ finden wir auch die zu (3) duale Eigenschaft:

(4) $A : \mathbb{N} \to \Sigma \implies \bigcap_{n\in\mathbb{N}} A(n) \in \Sigma$

(ii) Ist Σ eine σ-Algebra, dann nennen wir eine Abbildung der Form $\mu : \Sigma \to [0, \infty]$ ein **Maß** auf Σ, falls $\mu(\emptyset) = 0$ und für $A : \mathbb{N} \to \Sigma$ gilt:

$$\forall\, m \neq n \in \mathbb{N} : A(m) \cap A(n) = \emptyset \quad \Longrightarrow \quad \mu\left(\bigcup_{n\in\mathbb{N}} A(n)\right) = \sum_{n\in\mathbb{N}} \mu(A(n))$$

Man beachte, dass dabei die Summe unendlich werden kann, selbst wenn alle $\mu(A(n)) < \infty$ endlich sind. Ist μ ein Maß auf der σ-Algebra Σ, dann nennen wir (S, Σ, μ) auch einen **Maßraum**. Bezeichnet nun $\Omega(\mu)$ die Menge aller Mengen aus Σ mit endlichem Volumen $\mu(A) < \infty$, dann ist $\Omega(\Sigma)$ ein Mengenverband und $\mu : \Omega(\mu) \to \mathbb{R}^+$ ist ein Inhalt auf diesem Mengenverband:

$$\Omega(\mu) \;:=\; \{\, A \in \Sigma \mid \mu(A) < \infty \,\}$$

(iii) Ist (S, Σ, μ) ein Maßraum, dann nennen wir μ **endlich**, falls μ eine (und damit alle) der folgenden äquivalenten Eigenschaften erfüllt:

(a) $\mu(S) \;<\; \infty$

(b) $\Omega(\mu) \;=\; \Sigma$

(c) $\forall\, A \in \Sigma \;:\; \mu(A) < \infty$

Speziell im Fall $\mu(S) = 1$ sprechen wir auch von einem **Wahrscheinlichkeitsmaß** auf S. Schließlich nennen wir μ noch **σ-endlich**, falls es eine Folge messbarer Mengen $A : \mathbb{N} \to \Sigma$ gibt, mit

$$\forall\, n \in \mathbb{N} \;:\; \mu(A(n)) < \infty \quad \text{und} \quad S \;=\; \bigcup_{n \in \mathbb{N}} A(n)$$

(iv) Ist (S, Σ, μ) ein Maßraum, dann nennen wir eine beliebige Teilmenge $N \subseteq S$ eine **μ-Nullmenge** falls es eine Menge $A \in \Sigma$ gibt, mit $N \subseteq A$ und $\mu(A) = 0$. Und der Maßaum (S, Σ, μ) heißt **vollständig**, wenn jede μ-Nullmenge $N \subseteq S$ bereits zu $N \in \Sigma$ gehört, d.h. wenn für alle Teilmengen $N \subseteq S$ gilt:

$$\exists\, A \in \Sigma \;:\; N \subseteq A \text{ und } \mu(A) = 0 \;\Longrightarrow\; N \in \Sigma$$

Beweis:

- Wir zeigen zunächst, dass die σ-Algebra $\Sigma \subseteq \mathcal{P}(S)$ tatsächlich ein Mengenring ist. Zunächst ist $\emptyset = \complement S \in \Sigma$ aufgrund von Eigenschaften (1) und (2) von σ-Algebren. Und sind A, $B \in \Sigma$, dann definieren wir die Abbildung $H(0) := A$, $H(1) := B$ und für $n \geq 2$ noch $H(n) := \emptyset$. Dann ist $H : \mathbb{N} \to \Sigma$ wohldefiniert und offensichtlich gilt

$$A \cup B \;=\; A \cup B \cup \emptyset \cup \emptyset \cup \ldots \;=\; \bigcup_{n \in \mathbb{N}} H(n) \;\in\; \Sigma$$

- Weiter ist zu zeigen, dass beginnend mit einem Maß $\mu : \Sigma \to [0, \infty]$ die Sammlung von Teilmengen $\Omega(\mu)$ ein Mengenverband ist. Zunächst ist $\emptyset \in \Omega(\mu)$, da wir von Maßen ja $\mu(\emptyset) = 0 < \infty$ gefordert haben. Seien nun A, $B \in \Omega(\mu)$, d.h. es ist $\mu(A) < \infty$ und $\mu(B) < \infty$. In einem ersten Schritt betrachten wir $H(0) := B \setminus A$, $H(1) := A \cap B$ und für $n \geq 2$ auch $H(n) := \emptyset$. Da Σ auch ein Mengenring ist, ist damit $H : \mathbb{N} \to \Sigma$ wohldefiniert. Und wir haben eine disjunkte Zerlegung von

B gefunden:

$$B \;=\; (B \setminus A) \cup (A \cap B) \cup \emptyset \cup \emptyset \cup \ldots \;=\; \bigcup_{n \in \mathbb{N}} H(n)$$

Da μ ein Maß ist, können wir das Volumen von B also über die Summe der Volumen der $H(n)$ auswerten, und finden dabei:

$$\begin{aligned} \mu(B \setminus A) + \mu(A \cap B) \;&=\; \sum_{n \in \mathbb{N}} \mu(H(n)) \\ &=\; \mu\left(\bigcup_{n \in \mathbb{N}} H(n)\right) \;=\; \mu(B) \;<\; \infty \end{aligned}$$

Das bedeutet, dass die Volumina von $B \setminus A$ und $A \cap B$ beide endlich sein müssen und damit auch $B \setminus A$ und $A \cap B \in \Omega(\Sigma)$ sind.

Nun betrachten wir $H(0) := A \setminus B$, $H(1) := A \cap B$, $H(2) := B \setminus A$ und für $n \geq 3$ wieder $H(n) := \emptyset$. Wie oben ist $H : \mathbb{N} \to \Sigma$ damit wohldefiniert. Und offenbar ist $A \cup B$ damit disjunkt zerlegt worden:

$$A \cup B \;=\; (A \setminus B) \cup (A \cap B) \cup (B \setminus A) \cup \emptyset \cup \emptyset \cup \ldots \;=\; \bigcup_{n \in \mathbb{N}} H(n)$$

Wie gerade gesehen haben aber alle Bruchstücke $A \setminus B$, $A \cap B$ und $B \setminus A$ endliches Volumen unter μ und damit gilt auch

$$\begin{aligned} \mu(A \cup B) \;&=\; \mu\left(\bigcup_{n \in \mathbb{N}} H(n)\right) \;=\; \sum_{n \in \mathbb{N}} \mu(H(n)) \\ &=\; \mu(A \setminus B) + \mu(A \cap B) + \mu(B \setminus A) \;<\; \infty \end{aligned}$$

Es ist also auch $\mu(A \cup B) < \infty$ und damit $A \cup B \in \Omega(\Sigma)$. Insgesamt haben wir bewiesen, dass $\Omega(\Sigma)$ ein Mengenverband ist.

- Bleibt zu zeigen, dass μ ein Inhalt auf $\Omega(\Sigma)$ ist. Aber das ist trivial: sind A, $B \in \Omega(\mu)$ mit $A \cap B = \emptyset$, dann gilt für $H : \mathbb{N} \to \Sigma$ definiert, durch $H(0) := A$, $H(1) := B$ und für $n \geq 2$ wieder $H(n) := \emptyset$ offensichtlich:

$$\mu(A \cup B) \;=\; \mu\left(\bigcup_{n \in \mathbb{N}} H(n)\right) \;=\; \sum_{n \in \mathbb{N}} \mu(H(n)) \;=\; \mu(A) + \mu(B)$$

- Es ist klar, dass die Eigenschaften (b) und (c) endlicher Maße dieselbe Tatsache ausdrücken - nur anders geschrieben. Und (c) $\Longrightarrow$ (a) ist trivial. Es ist also noch (a) $\Longrightarrow$ (c) zu zeigen: sei also $A \in \Sigma$ beliebig, dann ist $\complement A \in \Sigma$ und $S = A \cup \complement A$ eine disjunkte Zerlegung von S. Da μ auch endlich additiv ist (wieder $H(0) := A, H(1) := \complement A$ und $H(n) := \emptyset$ für $n \geq 2$) erhalten wir: $\mu(A) + \mu(\complement A) = \mu(A \cup \complement A) = \mu(S) < \infty$. Und da die Summe endlich ist, kann keiner der Summanden unendlich sein und damit ist auch $\mu(A) < \infty$.

$\square$

Beispiel 11.2:

(i) Offensichtlich ist $\mathcal{P}(S) = \{ A \mid A \subseteq S \}$ auch eine σ-Algebra. Und es gibt auch immer ein Maß auf $\mathcal{P}(S)$, wenn es auch eher albern ist: $\mu(\emptyset) := 0$ und $\mu(A) := \infty$ für alle $\emptyset \neq A \subseteq S$. Ein schönes Beispiel, das nur zum Gegenbeispiel taugt.

(ii) Ein etwas interessanteres Maß ist das *Dirac-Maß*: Seien $S \neq \emptyset$ und $s \in S$ beliebig fixiert. Dann erhält man ein Wahrscheinlichkeitsmaß $\delta_s : \mathcal{P}(S) \to \{ 0, 1 \}$ durch $\delta_s(A) := 1$ falls $s \in A$ und $\delta_s(A) := 0$ falls $s \notin A$. Es ist dieses Maß, das ein Physiker meint, wenn er von der *Delta-Funktion* spricht: $\Delta_s : \mathcal{F}(S, \mathbb{R}) \to \mathbb{R} : f \mapsto f(s)$ ist eine lineare Abbildung, die man als Integral schreiben möchte. Mit Hilfe der Maßtheorie geht dies folgendermaßen (siehe auch Kapitel 17)

$$\Delta_s(f) \;=\; \int_S f(x)\, d\delta_s(x)$$

(iii) Ist $\Sigma \subseteq \mathcal{P}(S)$ eine σ-Algebra auf der Menge $S \neq \emptyset$, dann bilden die Maße auf Σ zwar keinen $\mathbb{R}$-Vektorraum, da wir ja nicht mit negativen Zahlen multiplizieren dürfen, aber einen positiven Kegel im Vektorraum der σ-additiven Funktionen $\sigma : \Sigma \to \mathbb{R}$. D.h. sind μ und $\nu : \Sigma \to [0, \infty]$ Maße und ist $a \geq 0$, dann erhält man ebenfalls Maße, durch $(a\mu)(A) := a \cdot \mu(A)$ und $(\mu + \nu)(A) := \mu(A) + \nu(A)$. Ein Speziallfall davon sind positive Linearkombinationen des Dirac-Maßes: sind $s(1), \ldots, s(n) \in S$ und $a(1), \ldots, a(n) \geq 0$, dann erhalten wir ein endliches Maß auf S durch

$$\sum_{i=1}^{n} a(i)\delta_{s(i)} \;:\; \mathcal{P}(S) \to \mathbb{R}^+ \;:\; A \mapsto \sum_{i=1}^{n} a(i) \cdot \delta_{s(i)}(A)$$

(iv) Ein weiteres einfaches Beispiel ist das *Zählmaß* auf einer endlichen Menge S. In dieser Fassung ist es sogar ein Wahrscheinlichkeitsmaß:

$$z \;:\; \mathcal{P}(S) \to [0, 1] \;:\; A \;\mapsto\; \frac{\#A}{\#S}$$

(v) Die Beispiele (iii) und (iv) lassen sich wie folgt verallgemeinern: Ist (S, p) ein Laplace-Raum, dann können wir diesen zu einem Wahrscheinlichkeitsraum $(S, \mathcal{P}(S), p)$ machen: Dabei wird p als Wahrscheinlichkeitsmaß wie zuvor schon definiert, als:

$$p \;:\; \mathcal{P}(S) \to [0, 1] \;:\; A \mapsto \sum_{s \in A} p(s)$$

(vi) Das wichtigste Beispiel haben wir bereits im vorangegangenen Kapitel in Beispiel (10.5.(v)) besprochen: das Lebesgue-Maß λ^d. Zunächst erinnern wir an die Bezeichnungen $\mathcal{I}^d = \left\{ [a, b[\mid a < b \in \mathbb{R}^d \right\}$ der halboffenen Quader und $\mathcal{O}^d = \left\{ U \subseteq \mathbb{R}^d \mid U \text{ offen} \right\}$ der offenen Teilmengen von $\mathbb{R}^d$.

Nach [Bauer, Maß und Integrationstheorie, 6.4] erzeugen beide Mengensysteme dieselbe σ-Algebra

$$\mathcal{B}(\mathbb{R}^d) \;:=\; \langle \mathcal{O}^d \rangle_\Sigma \;=\; \langle \mathcal{I}^d \rangle_\Sigma$$

Diese σ-Algebra wird als *Borel-σ-Algebra* von $\mathbb{R}^d$ bezeichnet. Das *Lebesgue-Maß* ist nun das eindeutig bestimmte Maß auf $\mathcal{B}(\mathbb{R}^d)$, mit

$$\lambda^d([a,b[) \;=\; \prod_{i=1}^{d}(b_i - a_i)$$

Dieses Maß lässt sich nach [Bauer, Maß und Integrationstheorie, 8.1] auch wie folgt charakterisieren: Ist $\mu : \mathcal{B}(\mathbb{R}^d)$ ein Maß auf der Borel-σ-Algebra von $\mathbb{R}^d$, so dass (1) für alle $A \in \mathcal{B}(\mathbb{R}^d)$ und alle $x \in \mathbb{R}^d$ gilt $\mu(x + A) = \mu(A)$ und (2) für den Einheitswürfel $I := [0,1] \subseteq \mathbb{R}^d$ gilt $m := \mu(I) < \infty$, dann ist bereits $\mu = m\lambda^d$. Und nach [Bauer, Maß und Integrationstheorie, 8.1] hat das Lebesgue-Maß dann folgende Transformationseigenschaft: Ist $T : \mathbb{R}^d \to \mathbb{R}^d$ linear und $A \in \mathcal{B}(\mathbb{R}^d)$ eine Borel-Menge, dann gilt für die Bildmenge $T(A)$:

$$\lambda^d(T(A)) \;=\; |\det T| \cdot \lambda^d(A)$$

(vii) Das Lebesgue-Maß kann (wie jedes andere Maß auch) dazu verwendet werden, weitere Maße zu konstruieren [Bauer, Maß und Integrationstheorie, 17.1]: Ist $f : \mathbb{R}^d \to [0,\infty]$ eine messbare Funktion (d.h. für alle $a \leq b \in \mathbb{R}$ gilt $f^{-1}([a,b]) \in \mathcal{B}(\mathbb{R}^d)$, das wird zum Beispiel von stetigen Funktionen erfüllt), dann erhalten wir ein weiteres Maß μ auf $\mathcal{B}(\mathbb{R}^d)$

$$\mu(A) \;:=\; \int_A f(x)\, d\lambda^d(x)$$

Für kompakte Quader $A = [a,b] \subseteq \mathbb{R}^d$ stimmt das hier verwendete Lebesgue-Integral übrigens mit dem Riemann-Integral überein [Bauer, Maß und Integrationstheorie, 16.4]. Doch auf die Konstruktion des Integrals werden wir erst in einem späteren Kapitel eingehen. Nach dem Satz von Radon-Nikodym ist die Existenz einer solchen Funktion f übrigens äquivalent, zu $\lambda^d(A) = 0 \implies \mu(A) = 0$ (für alle $A \in \mathcal{B}(\mathbb{R}^d)$).

(viii) Mit Blick auf die Normalverteilung möchten wir einen Spezialfall von (vii) geben: Sei $S \subseteq \mathbb{R}^d$ eine Borel-messbare Menge [d.h. $S \in \mathcal{B}(\mathbb{R}^d)$] und $\varrho : S \to \mathbb{R}^+$ eine positive, integrierbare Funktion, mit

$$\int_S \varrho(x)\, d\lambda^d(x) \;=\; 1$$

In diesem Fall nennen wir ϱ eine *Dichtefunktion* und wir erhalten einen Wahrscheinlichkeitsraum (S, Σ, p) durch Einschränkung von $\mathcal{B}(\mathbb{R}^p)$ auf S, d.h. $\Sigma = \mathcal{B}(\mathbb{R}^d) \cap S := \{\, B \cap S \mid B \in \mathcal{B}(\mathbb{R}^d) \,\}$ und

$$p \;:\; \Sigma \to [0,1] \;:\; A \mapsto \int_A \varrho(x) d\lambda^d(x)$$

Satz 11.3:

(i) Sei $S \neq \emptyset$ eine beliebige, nicht-leere Menge und zu jedem $i \in I$ (wobei auch $I \neq \emptyset$ beliebig) sei $\Sigma_i \subseteq \mathcal{P}(S)$ eine σ-Algebra auf S. Dann ist deren Schnitt wiederum eine σ-Algebra auf S:

$$\bigcap_{i \in I} \Sigma_i \quad \subseteq \quad \mathcal{P}(S) \quad \text{ist } \sigma\text{-Algebra}$$

Ist nun $\mathcal{U} \subseteq \mathcal{P}(S)$ irgend eine Sammlung von Teilmengen von S, dann können wir also die von $\mathcal{U}$ *erzeugte* σ-Algebra definieren, also die kleinste σ-Algebra, die $\mathcal{U}$ enthält:

$$\langle \mathcal{U} \rangle_\Sigma \quad := \quad \bigcap \{ \Sigma \subseteq \mathcal{P}(S) \mid \Sigma \text{ ist eine } \sigma\text{-Algebra, mit } \mathcal{U} \subseteq \Sigma \}$$

(ii) Ist $(S, \mathcal{A}, \mu)$ ein Raum mit Inhalt dann sind die folgenden Aussagen, betreffend μ, äquivalent [Bauer, Maß und Integrationstheorie, 3.2]:

(a) μ ist ein *Prämaß*, d.h. ist $A : \mathbb{N} \to \mathcal{A}$ eine Folge von Mengen aus $\mathcal{A}$, die paarweise disjunkt sind ($m \neq n \implies A(m) \cap A(n) = \emptyset$) so dass auch $\bigcup_n A(n) \in \mathcal{A}$ liegt, dann gilt:

$$\mu\left(\bigcup_{n \in \mathbb{N}} A(n)\right) \quad = \quad \sum_{n=0}^{\infty} \mu(A(n))$$

(b) μ ist *von unten stetig*, d.h. ist $A : \mathbb{N} \to \mathcal{A}$ eine aufsteigende Folge von Mengen aus $\mathcal{A}$ (also $A(0) \subseteq A(1) \subseteq A(2) \subseteq \dots$), so dass auch $\bigcup_n A(n) \in \mathcal{A}$ liegt, dann gilt:

$$\mu\left(\bigcup_{n \in \mathbb{N}} A(n)\right) \quad = \quad \lim_{n \to \infty} \mu(A(n))$$

(c) μ ist *von oben stetig*, d.h. ist $A : \mathbb{N} \to \mathcal{A}$ eine absteigende Folge von Mengen aus $\mathcal{A}$ (also $A(0) \supseteq A(1) \supseteq A(2) \supseteq \dots$), so dass auch $\bigcap_n A(n) \in \mathcal{A}$ liegt, dann gilt:

$$\mu\left(\bigcap_{n \in \mathbb{N}} A(n)\right) \quad = \quad \lim_{n \to \infty} \mu(A(n))$$

(d) μ ist $\emptyset$-*stetig*, d.h. ist $A : \mathbb{N} \to \mathcal{A}$ eine absteigende Folge von Mengen aus $\mathcal{A}$ (also wieder $A(0) \supseteq A(1) \supseteq A(2) \supseteq \dots$), so dass $\bigcap_n A(n) = \emptyset$ ist, dann gilt:

$$\lim_{n \to \infty} \mu(A(n)) \quad = \quad 0$$

(iii) Sei $(S, \mathcal{A}, \mu)$ ein Raum mit Inhalt, wobei μ sogar ein Prämaß ist (s. oben) und bezeichne $\Sigma = \langle \mathcal{A} \rangle_\Sigma$ die von $\mathcal{A}$ erzeugte σ-Algebra. Dann gibt es ein Maß $\overline{\mu} : \Sigma \to [0, \infty]$, das μ fortsetzt, d.h. so dass für alle $A \in \mathcal{A}$ gilt:

$$\overline{\mu}(A) \quad = \quad \mu(A)$$

Ist μ weiterhin σ-endlich, dann ist die Fortsetzung $\overline{\mu}$ sogar eindeutig bestimmt. Man gewinnt $\overline{\mu}$ durch Beschränkung des *äußeren Maßes* $\mu^* : \mathcal{P}(S) \to [0, \infty]$ von μ auf Σ: ist $B \subseteq S$ dann ist

$$\mu^*(B) \;:=\; \inf \left\{ \sum_{n \in \mathbb{N}} \mu(A(n)) \;\middle|\; A : \mathbb{N} \to \mathcal{A} \text{ mit } B \subseteq \bigcup_{n \in \mathbb{N}} A(n) \right\}$$

(Vorsicht: ein äußeres Maß ist kein Maß). Für einen Beweis dieser Aussage verweisen wir auf [Bauer, Maß und Integrationstheorie, 5.6].

(iv) Ist (S, Σ, μ) ein Maßraum, dann gibt es einen vollständigen Maßraum (S, T, ν) der (S, Σ, μ) fortsetzt, d.h. für den gilt $\Sigma \subseteq T$ und für alle Teilmengen $A \in \Sigma$ ist auch $\nu(A) = \mu(A)$. Man erhält eine solche Fortsetzung, durch Einschränkung des äußeren Maßes μ^* auf

$$T \;:=\; \{\, A \subseteq S \mid \forall\, B \subseteq S : \mu^*(B) \geq \mu^*(A \cap B) + \mu^*(B \setminus A) \,\}$$

Der kleinste vollständige Maßraum $(S, \widetilde{\Sigma}, \widetilde{\mu})$, der (S, Σ, μ) fortsetzt, wird die *Vervollständigung* von (S, Σ, μ) genannt. Das bedeutet $(S, \widetilde{\Sigma}, \widetilde{\mu})$ ist vollständig und ist (S, T, ν) ein weiter vollständiger Maßraum, mit $\Sigma \subseteq T$ und für alle $A \in \Sigma$ gilt $\nu(A) = \mu(A)$, dann gilt bereits $\widetilde{\Sigma} \subseteq T$ und für alle $A \in \widetilde{\Sigma}$ gilt auch $\widetilde{\mu}(A) = \nu(A)$.

(v) Ist $\mu : \mathcal{B}(\mathbb{R}) \to [0, \infty]$ ein Maß über den reellen Zahlen, dann definieren wir dessen **Verteilungsfunktion** F_μ durch:

$$F_\mu \;:\; \mathbb{R} \to [0, \infty] \;:\; b \mapsto \mu\,(]-\infty, b[)$$

Ist nun (S, Σ, p) ein Wahrscheinlichkeitsraum und $F = F_p$ die Verteilungsfunktion von p, dann besitzt F_p die folgenden Eigenschaften:

(1) F ist *monoton steigend*, d.h. sind $a \leq b \in \mathbb{R}$, dann gilt auch die Abschätzung $F(a) \leq F(b)$.

(2) F ist *linksseitig stetig*, d.h. für alle $b \in \mathbb{R}$ und alle $\varepsilon > 0$ gibt es ein $\delta > 0$ so dass für alle $a \in \mathbb{R}$ mit $0 < b - a < \delta$ gilt: $|F(b) - F(a)| < \varepsilon$.

(3) Für die linksseitigen bzw. rechtsseitigen Grenzwerte von $F(x)$ gilt:

$$\lim_{x \to -\infty} F(x) \;=\; 0 \quad \text{und} \quad \lim_{x \to \infty} F(x) \;=\; 1$$

Ist umkehrt $F : \mathbb{R} \to [0, \infty[$ eine Funktion mit den obigen Eigenschaften (1), (2) und (3), dann gibt es ein eindeutig bestimmtes Wahrscheinlichkeitsmaß $p : \mathcal{B}(\mathbb{R}) \to [0, 1]$ so dass $F = F_p$ die Verteilungsfunktion von p ist [Bauer, Maß und Integrationstheorie, 6.6]. Man erhält dieses durch Fortsetzung von

$$p([a, b[) \;=\; F(b) - F(a)$$

Definition 11.4: Zufallsvariablen:

(i) Seien S, T beliebige Mengen und $\mathcal{A} \subseteq \mathcal{P}(S)$ bzw. $\mathcal{B} \subseteq \mathcal{P}(T)$ zwei σ-Algebren darauf. Dann nennen wir eine Funktion $f : S \to T$ **messbar** (oder genauer **$\mathcal{A}$-$\mathcal{B}$-messbar**) falls für alle $B \in \mathcal{B}$ gilt: $f^{-1}(B) \in \mathcal{A}$. Und die Menge all dieser Abbildungen bezeichnen wir, mit

$$\mathcal{M}(S,T) \;:=\; \{\, f : S \to T \mid f \text{ ist messbar} \,\}$$

(ii) **Proposition:** Ist $\mathcal{V} \subseteq \mathcal{P}(T)$ eine Familie von Teilmengen von T und $\mathcal{B} = \langle \mathcal{V} \rangle_\Sigma$ die von $\mathcal{V}$ erzeugte σ-Algebra, dann genügt es die Messbarkeit von $f : S \to T$ nur für Mengen aus $\mathcal{V}$ zu überprüfen, denn:

$$f \text{ ist messbar} \iff \forall\, V \in \mathcal{V} \;:\; f^{-1}(V) \in \mathcal{A}$$

(iii) **Proposition:** Ist $f : S \to T$ beliebig, dann erhalten wir eine σ-Algebra auf S, durch $\Sigma(f) := \{\, f^{-1}(B) \mid B \in \mathcal{B} \,\}$. Offensichtlich ist f also immer $\Sigma(f)$-$\mathcal{B}$-messbar. Und mit dieser können wir die $\mathcal{A}$-$\mathcal{B}$-Messbarkeit von f offensichtlich formulieren, als

$$f \text{ ist messbar} \iff \Sigma(f) \;\subseteq\; \mathcal{A}$$

Ist $(S, \mathcal{A}, \mu)$ sogar ein Maßraum und $f : S \to T$ messbar, dann erhalten wir auch ein Maß (*push-forward* genannt) $f_*\mu$ auf T, durch

$$f_*\mu(B) \;:=\; \mu(f^{-1}(B))$$

(iv) Sei $I \neq \emptyset$ eine beliebige Indexmenge und zu jedem Index $i \in I$ sei eine σ-Algebra $\mathcal{B}_i \subseteq \mathcal{P}(T_i)$ auf der Menge T_i und eine Abbildung $f_i : S \to T_i$ gegeben. Dann definieren wir:

$$\Sigma(f_i \mid i \in I) \;:=\; \langle \bigcup_{i \in I} \{\, f^{-1}(B_i) \mid B_i \in \mathcal{B}_i \,\} \rangle_\Sigma$$

Proposition: Dies ist offenbar die kleinste σ-Algebra auf S, so dass alle f_i messbar werden. D.h. es gilt die Äquivalenz

$$\forall\, i \in I \;:\; f_i \text{ ist messbar} \iff \Sigma(f_i \mid i \in I) \;\subseteq\; \mathcal{A}$$

(v) Sei $\mathcal{A} \subseteq \mathcal{P}(S)$ immer noch eine σ-Algebra, dann fixieren wir nun $\mathcal{B} = \mathcal{B}(\mathbb{R})$ die Borel-σ-Algebra auf $\mathbb{R}$. Und wir nennen $X : S \to \mathbb{R}$ eine **Zufallsvariable** auf S, falls X messbar ist, d.h. falls gilt

$$\forall\, a < b \in \mathbb{R} \;:\; X^{-1}([a,b[) \;\in\; \mathcal{A}$$

Proposition: die Menge aller Zufallsvariablen auf S bezeichnen wir mit $\mathcal{Z}(S) := \mathcal{M}(S, \mathbb{R})$. Dann ist $\mathcal{Z}(S)$ eine kommutative $\mathbb{R}$-Algebra unter der punktweisen Addition und Multiplikation von Funktionen, die mit $X \in \mathcal{Z}(S)$ auch $X_+ = (|X| + X)/2$ und $X_- = (|X| - X)/2$ und $|X|$ enthält [Bauer, Maß und Integrationstheorie, 9.4 und 9.8].

(vi) Ist nun $\varphi : \mathbb{R} \to \{\,\text{wahr}, \text{falsch}\,\}$ ein Prädikat auf den reellen Zahlen, so dass $\varphi^{-1}(\text{wahr}) \in \mathcal{B}(\mathbb{R})$ ist - also zum Beispiel $\varphi(x) = (x = a)$ oder $\varphi(x) = (x > a)$ oder auch $\varphi(x) = (x^2 \leq a)$ für eine fest gewählte Zahl $a \in \mathbb{R}$ - dann verwenden wir auch die Schreibweise:

$$p(\varphi(X)) \;:=\; p\big(\,\{\, s \in S \mid \varphi(X(s))\,\}\,\big) \;\in\; [0,1]$$

(vii) Wir werden auf die Integration erst in Kapitel 17 eingehen, dennoch wollen wir hier schon den **Erwartungswert** $E(X)$, die **Varianz** $V(X)$ einer, bzw. die **Kovarianz** $\sigma(X \mid Y)$ zwischen zwei (hinreichend integrierbaren) Zufallsvariablen X und $Y \in \mathcal{Z}(S)$, auf dem Wahrscheinlichkeitsraum (S, Σ, p), definieren:

$$\begin{aligned}
E(X) &:= \int_S X(s)\,dp(s) \;=\; \int_{\mathbb{R}} x\,d(X_*p)(x) \\
\sigma(X \mid Y) &= E\left((X - E(X)) \cdot (Y - E(Y))\right) \\
&= E(XY) - E(X)E(Y) \\
V(X) &:= \sigma(X \mid X) \;=\; E\left((X - E(X))^2\right) \\
&= \int_S (X(s) - E(X))^2\,dp(s)
\end{aligned}$$

(viii) Ist $(S, \mathcal{A}, p)$ ein beliebiger Wahrscheinlichkeitsraum und $X \in \mathcal{Z}(S)$ eine Zufallsvariable mit Erwartungswert $\mu := E(X)$ und endlicher Varianz $\sigma^2 := V(X) < \infty$, dann definieren wir die **Standardisierung** von X:

$$Z \;:=\; \frac{X - \mu}{\sigma} \;\in\; \mathcal{Z}(S)$$

Aufgrund der Linearität des Erwartungswertes rechnet man leicht nach, dass Z nun zentriert ist und Varianz 1 besitzt:

$$\begin{aligned}
E(Z) &= \frac{1}{\sigma}(E(X) - \mu \cdot E(1)) \;=\; 0 \\
V(Z) &= E(Z^2) - E(Z)^2 \;=\; E(Z^2) \\
&= \frac{1}{\sigma^2} E\left((X - \mu)^2\right) \\
&= \frac{1}{\sigma^2} V(X) \;=\; 1
\end{aligned}$$

(ix) **Lemma:** *Tschebyscheff Ungleichung*: Die Varianz ist von zentraler Bedeutung, da sie eine Abschätzung für die Wahrscheinlichkeit der Abweichung vom Erwartungswert liefert. Sei (S, Σ, p) ein Wahrscheinlichkeitsraum und $X \in \mathcal{Z}(S)$ eine Zufallsvariable, mit $\mu := E(X)$ und $\sigma^2 := V(X) < \infty$. Sei ferner $\alpha > 1$ eine beliebige Zahl, dann gilt

$$p\left(|X - \mu| \geq \alpha\sigma\right) \;\leq\; \frac{1}{\alpha^2}$$

Beweis: Wie bezeichnen die Menge $A := \{ s \in S \mid |X(s) - \mu| \geq \alpha\sigma \}$, dann ist also $p(A) \leq 1/\alpha^2$ zu zeigen. Zunächst ist $A \in \mathcal{A}$, da X messbar ist. Wir bezeichnen mit $Z := (X - \mu)/\sigma$ wieder die Standardisierung von X. Dann ist offenbar $A = \{ s \in S \mid |Z(s)| \geq \alpha \} = \{ s \in S \mid Z(s)^2 \geq \alpha^2 \}$. Und nach Konstruktion von A gilt damit

$$\begin{aligned} \alpha^2 p(A) &= \int_A \alpha^2 dp(s) \leq \int_A Z(s)^2 dp(s) \\ &\leq \int_S Z(s)^2 dp(s) = E(Z^2) = 1 \end{aligned}$$

(x) Als Verallgemeinerung der Varianz in (vii) definiert man das n-te *Moment* der Zufallsvariablen $X \in \mathcal{Z}$ durch $m_n(X) := E(X^n)$ und bezeichnet $\mu := m_1(X) = E(X)$ den Erwartungswert, dann wird das n-te *zentrale Moment* von X definiert, als $z_n(X) := E((X - \mu)^n)$. In diesem Sinne ist die Varianz $V(X) = z_2(X)$ einfach das 2-te zentrale Moment. Die Berechnung der (zentralen) Momente kann sehr mühselig sein - deswegen bedient man sich gerne eines eleganten Tricks: Man betrachtet die *momente-erzeugende Funktion*

$$M_X(t) := E(e^{tX}) = \sum_{n=0}^{\infty} \frac{t^n}{n!} E(X^n) = \sum_{n=0}^{\infty} m_n(X) \frac{t^n}{n!}$$

In vielen Fällen kann man $M_X(t)$ effektiv berechnen und die Reihe konvergiert in einer Umgebung von $t = 0$. Dann aber kann man das n-te Moment durch n-malige Ableitung dieser Funktion gewinnen:

$$m_n(X) = (\partial_t^n M_X)(t = 0)$$

Bleibt die Frage wie man die zentralem Momente berechnen kann: sind die Momente bekannt, kann man die Binomialformel für $(X - \mu)^n$ ausnutzen. Aufgrund der Linearität von E findet man:

$$z_n(X) = \sum_{k=0}^{n} (-1)^{n-k} \binom{n}{k} \mu^{n-k} \cdot m_k(X)$$

(xi) Sei (S, Σ, p) ein Wahrscheinlichkeitsraum und $(X_n) \subseteq \mathcal{Z}(S)$ eine Folge von Zufallsvariablen auf S. Ist $X \in \mathcal{Z}(S)$ eine weiter Zufallsvariable auf S, dann sagen wir, dass (X_n) **p-fast sicher** gegen X **konvergiert**

$$X_n \to X \quad p\text{-fast sicher}$$

wenn eine der folgenden, äquivalenten Bedingungen erfüllt ist: siehe [Bauer, Maß und Integrationstheorie, 20.6] für einen Beweis

$$\forall \varepsilon > 0 \ : \ p\left(\bigcap_{n \in \mathbb{N}} \bigcup_{p \geq n} \{ s \in S \mid |X_p(s) - X(s)| > \varepsilon \} \right) = 0$$

$$\forall \varepsilon > 0 \ : \ \lim_{n \to \infty} p\left(\{ s \in S \mid \sup\{ |X_p(s) - X(s)| \mid p \geq n \} > \varepsilon \} \right) = 0$$

$$\forall\, \varepsilon > 0 \;:\; \lim_{n\to\infty} p\,(\{\, s \in S \;\mid\; \sup\{\, |X_p(s) - X(s)| \mid p \geq n \,\} < \varepsilon \,\}) \;=\; 1$$

(xii) Ist $X \in \mathcal{Z}(S)$ eine Zufallsvariable auf dem Wahrscheinlichkeitsraum (S, Σ, p) dann induziert X ein Wahrscheinlichkeitsmaß auf $\mathbb{R}$ durch push-forward $X_*p : \mathcal{B}(\mathbb{R}) \to [0,1] : B \mapsto p(X \in B)$. Wir kürzen die Sprechweise ab und nennen die Verteilungsfunktion dieses Maßes wiederum die **Verteilungsfunktion** $F_X : \mathbb{R} \to [0,1]$ von X selbst. D.h. für $b \in \mathbb{R}$ definieren wir

$$F_X(b) \;:=\; F_{X_*p}(b) \;=\; p\,(\{\, s \in S \mid X(s) < b \,\})$$

(xiii) Sei (S, Σ, p) ein Wahrscheinlichkeitsraum, $I \neq \emptyset$ wieder eine beliebige Indexmenge und zu jedem $i \in I$ sei $X_i \in \mathcal{Z}(S)$ eine Zufallsvariable auf S. Dann nennen wir die X_i **identisch verteilt**, falls alle X_i dieselbe Verteilungsfunktion auf $\mathbb{R}$ erzeugen, d.h. falls sie für alle i, $j \in I$ eine der folgenden vier, äquivalenten Bedingungen erfüllen:

(a) $F_{X_i} \;=\; F_{X_j}$

(b) $(X_i)_*p \;=\; (X_j)_*p$

(c) $\forall\, B \in \mathcal{B}(\mathbb{R}) \;:\; p(X_i^{-1}(B)) \;=\; p(X_j^{-1}(B))$

(d) $\forall\, a < b \in \mathbb{R} \;:\; p(X_i^{-1}([a,b[)) \;=\; p(X_j^{-1}([a,b[))$

Sind X_1 und $X_2 \in \mathcal{Z}(S)$ identisch verteilt, dann schreiben wir auch $X_1 \sim X_2$ und damit ist $\sim$ eine Äquivalenzrelation auf $\mathcal{Z}(S)$.

Beispiel 11.5:

(i) Für uns gibt es zwei wichtige Fälle: ist (S, p) ein Laplace Raum, dann betrachten wir wieder $(S, \mathcal{P}(S), p)$ den zugehörigen Wahrscheinlichkeitsraum. In diesem Fall ist offenbar jede Funktion $X : S \to \mathbb{R}$ eine Zufallsvariable $\mathcal{Z}(S) = \mathcal{F}(S, \mathbb{R})$. Und $X(S) \subseteq \mathbb{R}$ hat höchstens so viele Elemente wie S selbst, ist also abzählbar. Damit ist X genau dann integrierbar, wenn die Zahlenfamilie $s \mapsto X(s)p(s)$ summierbar ist und in diesem Fall wird der Erwartungswert zu

$$\begin{aligned} E(X) \;&=\; \int_S X(s)\, dp(s) \;=\; \sum_{s\in S} X(s)p(s) \\ &=\; \sum_{x\in X(S)} \sum_{X(s)=x} X(s)p(s) \;=\; \sum_{x\in X(S)} \sum_{X(s)=x} xp(s) \\ &=\; \sum_{x\in X(S)} x \sum_{X(s)=x} p(s) \;=\; \sum_{x\in X(S)} x\, p\left(X^{-1}(x)\right) \end{aligned}$$

Dabei schreibt man $p(X^{-1}(x)) = (X_*p)(x)$ oft auch suggestiv als $p(X = x)$ da es sich dabei um die Wahrscheinlichkeit handelt mit der X den Wert x annimmt. Insgesamt also

$$E(X) \;=\; \sum_{x\in X(S)} x\, p\,(X = x)$$

(ii) Der zweite wichtige Fall ist der einer *Dichtefunktion* $\varrho : S \to [0,1]$ auf einer Menge $S \in \mathcal{B}(\mathbb{R}^d)$. Wie im obigen Beispiel gesehen, erhalten wir daraus einen Wahrscheinlichkeitsraum $(S, \mathcal{B}(\mathbb{R}^d) \cap S, p)$ durch

$$p(A) \;:=\; \int_S \varrho(x)\, d\lambda^d(x)$$

Ist $X \in \mathcal{Z}(S)$ eine Zufallsvariable, dann ist X genau dann p-integrierbar, wenn ϱX über S λ^d-integrierbar ist. In diesem Fall wird der Erwartungswert von X, zu

$$E(X) \;=\; \int_S X(s)\, dp(s) \;=\; \int_S X(s)\varrho(s)\, d\lambda^d(s)$$

Definition 11.6: Produkträume

(i) Sei zu jedem $i \in 1 \ldots n$ ein σ-endlicher Maßraum (S_i, Σ_i, μ_i) gegeben. Dann erhalten wir einen weiteren σ-endlichen Maßraum - genannt den **Produktraum** der (S_i, Σ_i, μ_i) - durch

$$\bigotimes_{i=1}^{n} (S_i, \Sigma_i, \mu_i) \;:=\; (S, \Sigma, \mu)$$

Dabei ist der Grundraum S gerade das kartesische Produkt der S_i, d.h. die Menge aller n-Tupel deren i-ter Eintrag aus S_i stammt:

$$S \;:=\; \prod_{i=1}^{n} S_i \;=\; \{\, (s_1, \ldots, s_n) \mid \forall\, i \in 1 \ldots n \;:\; s_i \in S_i \,\}$$

Wir bezeichnen eine Menge der Gestalt $A_1 \times \cdots \times A_n$ wobei für alle $i \in 1 \ldots n$ jeweils $A_i \in \Sigma_i$ sei, als *Zylindermenge*. Die Menge aller Zylindermengen bezeichnen wir mit

$$\mathcal{C} \;:=\; \left\{ \prod_{i=1}^{n} A_i \;\middle|\; \forall\, i \in 1 \ldots n \;:\; A_i \in \Sigma_i \right\}$$

Offensichtlich ist $\mathcal{C} \subseteq \mathcal{P}(S)$ eine $\cap$-stabile Familie von Mengen auf S. Die von den Zylindermengen erzeugte σ-Algebra nehmen wir als

$$\Sigma \;:=\; \langle \mathcal{C} \rangle_\Sigma \;=\; \Sigma(\pi_i : S \to S_i \mid i \in 1 \ldots n)$$

Wir erhalten einen Inhalt μ auf den Zylindermengen $\mathcal{C}$ indem wir die Inhalte μ_i multiplizieren (siehe unten). Dieser Inhalt ist dann ein σ-endliches Prämaß und daher können wir μ dann zu einem Maß auf Σ fortsetzen. Wir bezeichnen die Fortsetzung wieder mit μ, dann ist μ also das eindeutig bestimmte Maß auf Σ mit der Eigenschaft

$$\forall \prod_{i=1}^{n} A_i \in \mathcal{C} \;:\; \mu\left(\prod_{i=1}^{n} A_i \right) \;=\; \prod_{i=1}^{n} \mu_i(A_i)$$

(ii) Sei nun $I \neq \emptyset$ eine beliebige Indexmenge und zu jedem $i \in I$ sei ein Wahrscheinlichkeitsraum (S_i, σ_i, p_i) gegeben. Dann erhalten wir einen weiteren Wahrscheinlichkeitsraum - genannt den **Produktraum** der (S_i, Σ_i, p_i) - durch

$$\bigotimes_{i \in I} (S_i, \Sigma_i, p_i) \quad := \quad (S, \Sigma, p)$$

Dabei ist der Grundraum S wieder das kartesische Produkt der S_i, d.h. die Menge aller I-Tupel deren i-ter Eintrag aus S_i stammt:

$$S \quad := \quad \prod_{i \in I} S_i \quad = \quad \{\, (s_i) \mid \forall\, i \in I \;:\; s_i \in S_i \,\}$$

Die Menge aller nicht-leeren, endlichen Teilmengen von I bezeichnen wir wieder mit $\Omega(I) := \{\, \alpha \subseteq I \mid \alpha \neq \emptyset \text{ und } \#\alpha < \infty \,\}$. Damit definieren wir dann wieder die Menge der Zylindermengen von S, vermöge

$$\mathcal{C} \;:=\; \bigcup_{\alpha \in \Omega(I)} \mathcal{C}_\alpha \quad \text{wobei} \quad \mathcal{C}_\alpha \;:=\; \left\{ \prod_{i \in I} A_i \;\middle|\; \begin{array}{l} \forall\, i \in I : A_i \in \Sigma_i \\ \forall\, i \notin \alpha : A_i = S_i \end{array} \right\}$$

D.h. eine Zylindermenge ist in diesem Fall wieder ein kartesisches Produkt von Mengen $A_i \in \Sigma_i$ es sind aber nur endlich viele Mengen $A_i \neq S_i$ zugelassen. Ist I endlich, so ist $\mathcal{C}$ also genau das oben definerte System von Zylindermengen. Und wieder ist $\mathcal{C}$ dann $\cap$-stabil und die von den Zylindermengen erzeugte σ-Algebra nehmen wir als

$$\Sigma \quad := \quad \langle \mathcal{C} \rangle_\Sigma \quad = \quad \Sigma(\pi_i : S \to S_i \mid i \in I)$$

Man beachte, dass $p_i(S_i) = 1$ keinen Beitrag zu dem Produkt leistet, und deshalb erhalten wir wieder ein Wahrscheinlichkeitsmaß p auf Σ, dass eindeutig bestimmt wird, durch die Eigenschaft

$$\forall\, \alpha \in \Omega(I)\; \forall \prod_{i=1}^{n} A_i \in \mathcal{C}_\alpha \quad : \quad p\left(\prod_{i \in I} A_i \right) \quad = \quad \prod_{i \in \alpha} p_i(A_i)$$

Beispiel 11.7:
Sei I eine abzählbare (zum Beispiel endliche) Menge und zu jedem $i \in I$ sei ein Laplace-Raum (S_i, p_i) gegeben. Dann stammt der Produktraum der zugehörigen Wahrscheinlichkeitsräume $(S_i, \mathcal{P}(S_i), p_i)$ wiederum von einem Laplace-Raum (S, p) und dieser ist gegeben, durch:

$$\begin{aligned} \bigotimes_{i \in I} (S_i, p_i) \quad &= \quad (S, p) \\ S \quad &= \quad \prod_{i \in I} S_i \\ p((s_i)) \quad &= \quad \prod_{i \in I} p_i(s_i) \end{aligned}$$

Kapitel 12

Unabhängigkeit

In Kapitel 7 haben wir den Begriff der *bedingten Wahrscheinlichkeit* $p(A \mid B)$ des Eintreffens eines Ereignisses A unter der Bedingung B definiert. Wenn sich diese Wahrscheinlichkeit aber gar nicht durch die Bedingung B verändert, also wenn $p(A \mid B) = p(A)$ gilt, dann gibt es Sinn zu sagen, dass A von B *unabhängig* ist. Und wir haben gesehen, dass eine äquivalente Formulierung dieser Eigenschaft lautet

$$p(A \cap B) \;=\; p(A) \cdot p(B)$$

Wir nehmen diese Gleichung als Grundlage für die Definition der Unabhängigkeit einer Menge von Ereignissen. Die Unabhängigkeit einer Menge von Ereignissen wird dann erweitert zur Unabhängigkeit von Zufallsvariablen. Wir werden feststellen, dass der Erwartungswert bei unabhängigen Zufallsvariablen multiplikativ und die Varianz bei unabhängigen Zufallsvariablen additiv verhält. Doch das eigentliche Ziel dieses Kapitels ist das Gesetz der großen Zahlen, das besagt, dass der Durchschnitt bei unabhängiger Wiederholung eines Zufallsexperimentes tatsächlich gegen den Erwartungswert konvergiert. Dieser Satz bietet also den Übergang zwischen Statistik und Stochastik.

Definition 12.1: Unabhängigkeit:

(i) Sei (S, Σ, p) ein Wahrscheinlichkeitsraum und $\mathcal{U} \subseteq \Sigma$ eine beliebige Menge von Ereignissen aus S. Dann nennen wir $\mathcal{U}$ **unabhängig** falls für je endlich viele $(1 \leq n \in \mathbb{N})$ Ereignisse $A_1, \ldots, A_n \in \mathcal{U}$ gilt

$$p(A_1 \cap \cdots \cap A_n) \;=\; p(A_1) \cdot \cdots \cdot p(A_n)$$

(ii) **Satz:** *Lemma von Borel-Cantelli* [Bauer, Wahrscheinlichkeitstheorie 11.1]: Ist $\{\, A_n \mid n \in \mathbb{N} \,\} \subseteq \Sigma$ eine Folge von Ereignissen, für die die Reihe über $p(A_n)$ konvergiert, dann ist die die Wahrscheinlichkeit, dass unendlich viele dieser Ereignisse gleichzeitig eintreten 0

$$\sum_{n=0}^{\infty} p(A_n) < \infty \quad \Longrightarrow \quad \lim_{n \to \infty} p\left(\bigcup_{k=n}^{\infty} A_k \right) \;=\; 0$$

Sind die A_n paarweise unabhängig (d.h. für alle $m \neq n \in \mathbb{N}$ sind A_m und A_n unabhängig) und divergiert die Reihe über die $p(A_n)$, dann ist die die Wahrscheinlichkeit, dass unendlich viele dieser Ereignisse gleichzeitig eintreten 1

$$\sum_{n=0}^{\infty} p(A_n) = \infty \quad \Longrightarrow \quad \lim_{n\to\infty} p\left(\bigcup_{k=n}^{\infty} A_k\right) \;=\; 1$$

(iii) Sei $I \neq \emptyset$ eine beliebige, nicht-leere Indexmenge und zu jedem $i \in I$ sei eine Menge $\mathcal{U}_i \subseteq \Sigma$ von Ereignissen gegeben. Dann heißt die Familie $(\mathcal{U}_i)$ **unabhängig**, falls für je endlich viele Indices $\{\, i(1), \ldots, i(n) \,\} \subseteq I$ und Mengen $A_k \in \mathcal{U}_{i(k)}$ gilt:

$$p(A_1 \cap \cdots \cap A_n) \;=\; p(A_1) \cdot \cdots \cdot p(A_n)$$

(iv) **Satz:** [Bauer, Wahrscheinlichkeitstheorie 6.4]: Ist für alle $i \in I$ die Menge $\mathcal{U}_i \subseteq \Sigma$ $\cap$-stabil und bezeichnet $\Sigma_i := \langle \mathcal{U}_i \rangle_\Sigma \subseteq \Sigma$ die von $\mathcal{U}_i$ erzeugte σ-Algebra, dann sind die folgenden beiden Aussagen äquivalent:

(a) $(\mathcal{U}_i)$ ist unabhängig

(b) (Σ_i) ist unabhängig

(v) Sei wieder $I \neq \emptyset$ eine beliebige, nicht-leere Indexmenge und zu jedem $i \in I$ sei eine messbare Funktion $f_i : S \to T_i$ für irgendeine Menge T_i mit einer σ-Algebra $\mathcal{B}_i \subseteq \mathcal{P}(T_i)$ gegeben. Dann nennen wir die Familie (f_i) **unabhängig** falls die Familie der σ-Algebren $(\Sigma(f_i))$ unabhängig ist. Anders gesagt, falls für je endlich viele Indices $\{\, i(1), \ldots, i(n) \,\} \subseteq I$ und Mengen $B_k \in \mathcal{B}_{i(k)}$ gilt:

$$p\left(\bigcap_{k=1}^{n} f_{i(k)}^{-1}(B_k)\right) \;=\; \prod_{k=1}^{n} p\left(f_{i(k)}^{-1}(B_k)\right)$$

(vi) **Satz:** [Bauer, Wahrscheinlichkeitstheorie 8.1]: Sei (S, Σ, p) ein Wahrscheinlichkeitsraum und $X_1, \ldots, X_n \in \mathcal{Z}(S)$ seien integrierbare Zufallsvariablen auf S. Sind diese unabhängig (d.h. die Familie (X_i) ist unabhängig), dann gilt:

$$E\left(\prod_{i=1}^{n} X_i\right) \;=\; \prod_{i=1}^{n} E(X_i)$$

(vii) **Satz**: [Bauer, Wahrscheinlichkeitstheorie 8.3]: Sei (S, Σ, p) ein Wahrscheinlichkeitsraum und $X_1, \ldots, X_n \in \mathcal{Z}(S)$ seien integrierbare Zufallsvariablen auf S. Dann rechnet man leicht nach, dass:

$$V\left(\sum_{i=1}^{n} X_i\right) \;=\; \sum_{i=1}^{n} V(X_i) + 2 \sum_{1 \leq i < j \leq n} \sigma(X_i \mid X_j)$$

Sind die X_i paarweise unabhängig, so gilt insbesondere $\sigma(X_i \mid X_j) = 0$ wegen (vi) (man sagt die X_i sind auch *unkorreliert*). Und damit ist die Varianz additiv auf paarweise unabhängigen Zufallsvariablen:

$$V\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} V(X_i)$$

(viii) **Satz:** *Chernoff-Schranken*: Ist (S, Σ, p) ein Wahrscheinlichkeitsraum und sind X_1 bis $X_n \in \mathcal{Z}(S)$ unabhängige Zufallsvariablen auf S, mit Wertebereich $X_i(S) \subseteq [0,1]$ für alle $i \in 1 \ldots n$ gilt, dann bezeichnen wir $X := X_1 + \ldots + X_n \in \mathcal{Z}(S)$ und $\mu := E(X) = E(X_1) + \ldots + E(X_n)$. Und damit gilt dann für alle $\alpha \geq 0$

$$\begin{aligned} p(X \geq (1+\alpha)\mu) &\leq \exp\left(-\frac{\alpha^2}{2+\alpha}\mu\right) \\ p(X \geq (1-\alpha)\mu) &\leq \exp\left(-\frac{\alpha^2}{2}\mu\right) \\ p(|X-\mu| \geq \alpha\mu) &\leq 2 \cdot \exp\left(-\frac{\alpha^2}{2+\alpha}\mu\right) \end{aligned}$$

(ix) **Satz:** Sind X und $Y \in \mathcal{Z}(S)$ unabhängige Zufallsvariablen auf dem Wahrscheinlichkeitsraum (S, Σ, p), mit den Dichtefunktionen f_X bzw. f_Y d.h. für alle $b \in \mathbb{R}$ gilt

$$F_X(b) = \int_{-\infty}^{b} f_X(t)dt \quad \text{und } F_Y(b) = \int_{-\infty}^{b} f_Y(t)dt$$

dann hat die Summe $X + Y \in \mathcal{Z}(S)$ die Dichtefunktion $f_X * f_Y$ der Faltung von f_X und f_Y. D.h. für alle $b \in \mathbb{R}$ gilt explizit

$$F_{X+Y}(b) = \int_{-\infty}^{b} \int_{\mathbb{R}} f_X(t) \cdot f_Y(x-t)dt\,dx$$

Insbesondere: sind X und $Y : S \to \mathbb{Z}$ diskrete Zufallsvariablen dann gilt für die Wahrscheinlichkeit, dass $X + Y = k \in \mathbb{Z}$ ist:

$$p(X+Y=k) = \sum_{i \in \mathbb{Z}} p(X=i) \cdot p(Y=k-i)$$

Beispiel 12.2:

(i) Wir werfen zwei Würfel, d.h. wir betrachten den Wahrscheinlichkeitsraum $(S, \mathcal{P}(S), z)$, wobei $S = (1 \ldots 6)^2$ und z das Zählmaß (siehe oben) ist. Wir betrachten nun die folgenden Ereignisse

$$\begin{aligned} U_1 &:= \{\, (d_1, d_2) \in S \mid d_1 \text{ ungerade} \,\} \\ U_2 &:= \{\, (d_1, d_2) \in S \mid d_2 \text{ ungerade} \,\} \\ V &:= \{\, (d_1, d_2) \in S \mid d_1 + d_2 \text{ ungerade} \,\} \end{aligned}$$

Offensichtlich sind alle Paare $\{U_1, U_2\}$, $\{U_1, V\}$ und $\{U_2, V\}$ unabhängig. Aber $\{U_1, U_2, V\}$ ist *nicht unabhängig*, da $U_1 \cap U_2 \cap V = \emptyset$ unmöglich ist. Es ist also nicht möglich von der paarweisen Unabhängigkeit auf die gesamte Unabhängigkeit zu schließen.

(ii) Wir betrachten wieder den Raum von 3 Würfeln, d.h. diesmal sei $S = (1 \ldots 6)^3$. Mit $X_k : S \to 1 \ldots 6 \subseteq \mathbb{R}$ wählen wir das Ergebnis des k-ten Würfels aus, d.h. $X_k(d_1, d_2, d_3) := d_k$. Dann sind X_1, X_2 und X_3 voneinander unabhängig und identisch verteilt. Betrachtet man hingegen die Summen $S_L = X_1 + X_2$ der ersten, bzw. $S_R = X_2 + X_3$ der letzten beiden Würfel [also S_L und $S_R : S \to 2 \ldots 12 \subseteq \mathbb{R}$ mit $S_L(d_1, d_2, d_3) := d_1 + d_2$ beziehungsweise $S_R(d_1, d_2, d_3) = d_2 + d_3$] dann sind S_L und S_R wieder identisch verteilt, aber nicht mehr unabhängig voneinander.

(iii) Wir betrachten den Raum von n Münzwürfen: $S = \{0, 1\}^n$. Wieder wird $(S, \mathcal{P}(S), z)$ zum Wahrscheinlichkeitsraum, wobei $\#S = 2^n$ Elemente hat. Dann sind zwei binäre Zufallsvariablen X, $Y : S \to \{0, 1\} \subseteq \mathbb{R}$ genau dann gleich verteilt, wenn sie gleich oft, das Ergebnis 1 produzieren, d.h. wenn gilt:

$$\#\{s \in S \mid X(s) = 1\} \quad = \#\{s \in S \mid Y(s) = 1\}$$

Es kommt dabei nicht darauf an, *welche* der Wurffolgen von X bzw. Y ausgewählt werden - es genügt, dass es gleich viele sind. Insbesondere können X und Y unabhängig sein, oder auch nicht.

(iii) Ist allgemein zu jedem $i \in I$ (I eine beliebige Indexmenge) ein Wahrscheinlichkeitsraum (S_i, Σ_i, p) gegeben und bezeichnen wir deren Produktraum wieder mit (S, Σ, p) dann sind die *kanonischen Projetionen*

$$\pi_k \ : \ S \to S_i \ : \ (s_i) \mapsto s_k$$

nach Konstruktion unabhängige Zufallsvariablen. D.h. der Produktraum modelliert die unabhängige Durchführung der I verschiedenen Zufallsexperimente (S_i, Σ_i, p_i). Sind diese Räume identisch, dann wäre das also die I-malige Ausführung desselben Zufallsexperiments. Beispiele wären n-maliges Würfeln oder Ziehen mit Zurücklegen.

(iv) Ist (S, Σ, p) ein beliebiger Wahrscheinlichkeitsraum und A_1 bis A_n eine endliche Überdeckung von S, d.h. es gilt $S = A_1 \cup \cdots \cup A_n$. Sind diese Ereignisse A_1 bis A_n unabhängig, dann muss es zwingend ein fastsicheres Ereignis A_k darunter geben. D.h. es gilt

$$\{A_1, \ldots, A_n\} \text{ unabhängig} \quad \Longrightarrow \quad \exists\, k \in 1 \ldots n \ : \ p(A_k) \ = \ 1$$

Denn die Unabhängigkeit der A_i besagt gerade, dass für alle $I \subseteq 1 \ldots n$ gilt, dass $p(\bigcap_{i \in I} A_i) = \prod_{i \in I} p(A_i)$ ist. Wegen $p(S) = 1$ ist das aber gerade die zweite Voraussetzung in (10.6), so dass $p(A_k) = p(S)$ folgt.

Satz 12.3: von Etemadi: Gesetz der großen Zahlen
Sei (S, Σ, p) ein Wahrscheinlichkeitsraum und $X_n \in \mathcal{Z}(S)$ (mit $n \in \mathbb{N}$) eine Folge von Zufallsvariablen auf S, so dass gilt:

(1) Alle X_n sind p-integrierbar (d.h. es ist $\int X_n \, dp < \infty$, näheres dazu findet sich im Kapitel 17 zur Integrationstheorie),

(2) Die X_n sind identisch verteilt, d.h. es gibt eine Funktion $F : \mathbb{R} \to [0, 1]$, so dass für alle $n \in \mathbb{N}$ und alle $x \in \mathbb{R}$ gilt $p(X_n \leq x) = F(x)$,

(3) Die X_n sind paarweise unabhängig, d.h. für alle $m \neq n \in \mathbb{N}$ und alle Borel-Mengen B_m, $B_n \in \mathcal{B}(\mathbb{R})$ gilt

$$p\Big(X_m^{-1}(B_m) \cap X_n^{-1}(B_n)\Big) \;=\; p\big(X_m^{-1}(B_m)\big) \cdot p\big(X_n^{-1}(B_n)\big)$$

Dann konvergiert für $n \to \infty$ der Durchschnitt der X_n p-fast sicher gegen den Durchschnitt der Erwartungswerte $E(X_n)$, formal ausgedrückt:

$$\frac{1}{n} \sum_{k=1}^{n} (X_k - E(X_k)) \;\to\; 0 \quad p\text{-fast sicher}$$

Bemerkung 12.4:
Ist (S, p) ein Laplace Raum, dann betrachten wir den zugehörigen Wahrscheinlichkeitsraum $(S, \mathcal{P}(S), p)$ und zu $s \in S$ bezeichne $X_s : S \to \{\, 0, 1 \,\} \subseteq \mathbb{R}$ die Funktion $X_s(t) := 1$ für $t = s$ und $X_s(t) = 0$ sonst, die testet, ob s vorliegt. Wir können das Experiment (ziehen und schauen ob wir s erwischt haben) beliebig oft wiederholen. Starten wir unter immer gleichen Ausgangsbedingungen (also *mit* Zurücklegen), so sind die einzelnen Versuche voneinander unabhängig und der Raum der n-fachen Durchführung modelliert, durch

$$(S, \mathcal{P}(S), p)^{\otimes n} \;=\; (S^n, \mathcal{P}(S^n), p^n)$$

Um eine Grenzwertbetrachtung durchführen zu können müssen wir unendlich viele Durchführungen des Experiments zulassen, d.h. wir müssen unendlich lange Folgen $(s_k) \in S^{\mathbb{N}}$ betrachten. Der zugehörige Raum ist

$$(S, \mathcal{P}(S), p)^{\otimes \mathbb{N}} \;=\; (S^{\mathbb{N}}, \mathcal{P}(S^{\mathbb{N}}), p^{\mathbb{N}})$$

Wir bezeichnen mit $\pi_k : S^{\mathbb{N}} \to S : (s_i) \mapsto s_k$ wieder die kanonischen Projektionen und mit $X_k := X_s \pi_k : S^{\mathbb{N}} \to \{\, 0, 1 \,\}$ den Test auf das Vorkommen von s an k-ter Stelle. Offensichtlich sind die X_k unabhängig und identisch verteilt, so dass gilt

$$\frac{1}{n} \sum_{k=1}^{n} (X_k - E(X_k)) \;\to\; 0 \quad p\text{-fast sicher}$$

Es ist $E(X_k) = 1 \cdot p(X_k = s) = p(s)$ konstant und damit können wir die $E(X_k)$ aus der Summe ziehen und auf die andere Seite stellen, zu

$$\frac{1}{n}\sum_{k=1}^{n} X_k \;\to\; p(s) \quad p\text{-fast sicher}$$

D.h. die (mit p gemessene) relative Häufigkeit von s bei n Ziehungen nähert sich der zugrunde gelegten Wahrscheinlichkeit $p(s)$ an. Wir haben also eine exakte Formulierung des empirischen *Gesetzes der großen Zahlen* gefunden.

Aus dem Gesetz der großen Zahlen erhält man ein weiteres starkes Resultat, dass nicht umsonst auch als *Fundamentalsatz der mathematischen Statistik* bezeichnet wird:

Korollar 12.5: von Glivenko-Cantelli:
Sei (S, Σ, p) ein Wahrscheinlichkeitsraum und $X_n \in \mathcal{Z}(S)$ (mit $n \in \mathbb{N}$) wieder eine Folge von Zufallsvariablen auf S, so dass gilt:

(1) Alle X_n sind p-integrierbar (d.h. es ist $\int X_n \, dp < \infty$, näheres dazu findet sich im Kapitel 17 zur Integrationstheorie),

(2) Die X_n sind identisch verteilt, d.h. es gibt eine Funktion $F : \mathbb{R} \to [0,1]$, so dass für alle $n \in \mathbb{N}$ und alle $x \in \mathbb{R}$ gilt $p(X_n \leq x) = F(x)$,

(3) Die X_n sind paarweise unabhängig, d.h. für alle $m \neq n \in \mathbb{N}$ und alle Borel-Mengen B_m, $B_n \in \mathcal{B}(\mathbb{R})$ gilt

$$p\Big(X_m^{-1}(B_m) \cap X_n^{-1}(B_n)\Big) \;=\; p\big(X_m^{-1}(B_m)\big) \cdot p\big(X_n^{-1}(B_n)\big)$$

Dann definieren wir die *empirische Verteilungsfunktion* $F_n : S \times \mathbb{R} \to [0,1]$ der ersten n Zufallsvariablen X_1 bis X_n als den Anteil

$$F_n(s)(x) \;:=\; \frac{\#\{\, i \in 1 \ldots n \mid X_i(s) \leq x \,\}}{n}$$

Dann konvergiert $x \mapsto F_n(s)(x)$ für p-fast alle gleichmäßig gegen die gemeinsame Verteilungsfunktion F der Zufallsvariablen X_n, d.h. es gilt

$$p\left(\left\{ s \in S \;\middle|\; \lim_{n\to\infty} \|F - F_n(s)\|_\infty \right\} = 0\right) \;=\; 1$$

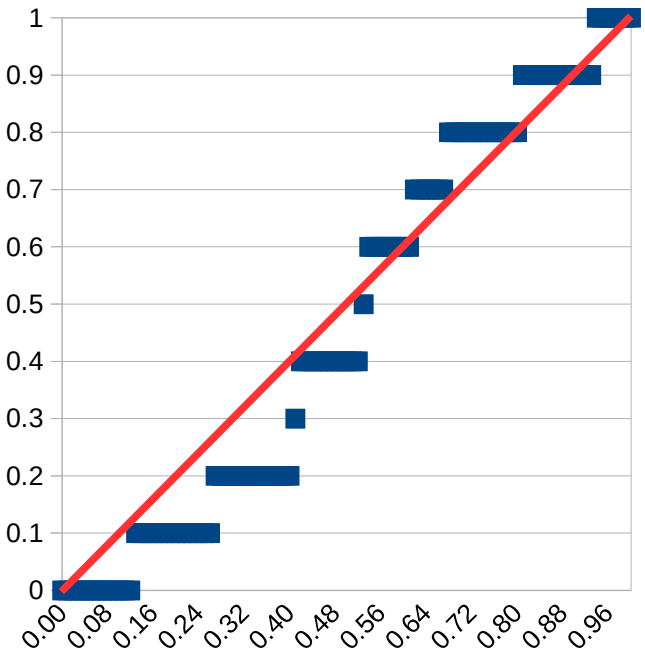
Bemerkung: von *gleichmäßiger Konvergenz* der Folge $(F_n(s))$ von Funktionen gegen die Funktion F spricht man, wenn der maximale Abstand (gemessen durch das Supremum) der beiden Funktionen im Limes gegen 0 geht. Dieser Abstand wird gemessen durch die *Supremumsnorm*

$$\|F - F_n(s)\|_\infty \;=\; \sup_{x\in\mathbb{R}} |F(x) - F_n(s)(x)|$$

Beispiel 12.6:
Die folgende Abbildung illustriert die Situation bei einer Gleichverteilung auf $S = [0, 1]$, d.h. es ist $p([a, b]) = b - a$ für $0 \leq a \leq b \leq 1$. D.h. wir ziehen eine zufällige Zahl $x \in S$, wobei alle Zahlen gleich wahrscheinlich sind. Um beliebig viele unabhängige Ziehungen zu modellieren, verwenden wir wieder den Produktraum

$$(S, \mathcal{B}(\mathbb{R}) \cap S, p)^{\otimes \mathbb{N}}$$

Die n-te Ziehung aus S bezeichnen wir mit X_n, d.h. es ist $X_n : S^{\mathbb{N}} \to S : (x_i) \mapsto x_n$. Die Verteilungsfunktion von X_n ist damit immer $F(x) = x$ (für $x \in S$, für $x \leq 0$ ist $F(x) = 0$ und für $x \geq 1$ ist $F(x) = 1$). In der folgenden Abbildung haben wir $n = 10$ dargestellt: d.h. wir haben 10 Zufallszahlen $x_i \in S$ bestimmt, deren (empirische) Verteilungsfunktion ist blau eingezeichnet. Man sieht, dass diese in der Nähe der rot gezeichneten (theoretischen) Verteilungsfunktion F liegt. Je größer n wird (je mehr Zahlen wir ziehen), desto besser wird die Annäherung. Der obige Satz sagt, dass die Konvergenz sogar (fast sicher) gleichmäßig ist:



Vorsicht: die empirische Verteilungsfunktion $F_n(S)$ in obigem Satz wurde über die Zufallsvariablen X_n definiert. Sie *modelliert* zwar die n-fache Ziehung aus S, das sagt aber nicht viel über eine tatsächliche, *real* erfolgende Ziehung. $F_n(s)$ ist nur der Durchschnitt über (die Verteilungsfunktionen) *aller* möglichen Ziehungen. Und wie immer sagt eine Statistik leider nichts über den Einzelfall aus. In diesem Fall gibt es andere Zufallszahlen x_i, deren Verteilungsfunktion eine deutlich schlechtere Übereinstimmung mit der Diagonalen zeigt, aber eben auch solche bei denen die Übereinstimmung besser ist. Wir *wissen* also nur, dass die durchschnittliche Abweichung im Grenzwert $n \to \infty$ fast sicher 0 wird.

Kapitel 13

Normalverteilung

Normalverteilung:
Beginnen wir mit einem einfachen Experiment: wir werfen 4 (6-seitige) Würfel und addieren deren Augensummen. Das Ergebnis wird also eine Zahl von 4 bis 24 sein, doch die Ergebnisse werden nicht gleich häufig sein: um die Zahl 4 zu erhalten, müssen alle 4 Würfel eine 1 zeigen, um eine 14 zu erreichen gibt es sehr sehr viele Möglichkeiten. Wir beobachten also eine Konzentration auf mittlere Werte, da die einzelnen Würfelergebnisse sich ja gegenseitig ausgleichen können.

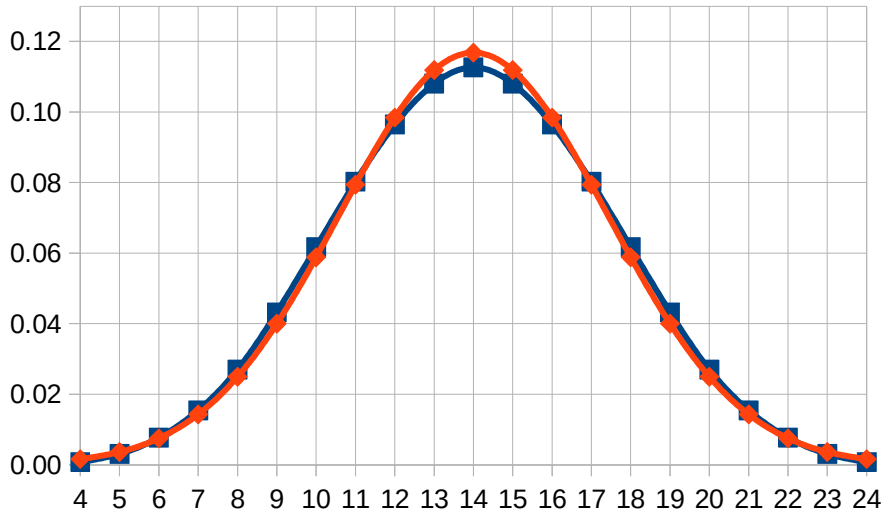
Bei $m \geq 4$ Würfeln, wird dieser Effekt noch deutlicher - die Randereignisse m und $6m$ werden im Vergleich zum Erwartungswert $3.5 \cdot m$ immer seltener. Das beobachtet man auch beim Galton-Brett [eine einfache Simulation dessen findet man unter http://www.heise.de/download/galton-brett-1116390.html]. Je mehr Stufen man im Galton-Brett durchlaufen lässt, desto besser nähert sich die Häufigkeitsverteilung einer Grenzverteilung - einer sog. *Normalverteilung* - an. Das erstaunliche ist, dass sowohl das Würfelexperiment, als auch das Galtonbrett denselben Typ Grenzverteilung besitzen.

Doch eigentlich ist dies gar nicht so erstaunlich: In beiden Fällen wird ein und derselbe Versuch immer wieder (unabhängig voneinander) wiederholt, so dass sich die einzelnen Ergebnisse gegenseitig aufheben können, wodurch eine Konzentration in der Mitte entsteht. Dies ist gerade die Aussage des zentralen Grenzwertsatzes: Kombiniert man m Versuche mit derselben Wahrscheinlichkeitsverteilung so, dass die einzelnen Versuche voneinander unabhängig sind, dann nähert sich das Ergebnis im Grenzwert $m \to \infty$ einer Normalverteilung an. Eine exakte Formulierung dieses Ergebnisses werden wir noch in diesem Kapitel geben.

Doch zunächst zurück zu unserem Ausgangsexperiment: der Augensumme von 4 Würfeln. Eine einfache kombinatorische Überlegung (Induktion über m) liefert die Anzahl der Kombinationen die die Augensumme $s \in S = 4 \ldots 24$ liefern. Und da es insgesamt $6^4 = 1296$ Kombinationen gibt, erhält man daraus die folgende Häufigkeitsverteilung der Ergebnisse:

s	$n(s)$	$h(s)$	$g(\mu,\sigma,s)$	s	$n(s)$	$h(s)$	$g(\mu,\sigma,s)$
4	1	0.0008	0.0016	15	140	0.1080	0.1119
5	4	0.0031	0.0036	16	125	0.0965	0.0984
6	10	0.0077	0.0075	17	104	0.0802	0.0794
7	20	0.0154	0.0143	18	80	0.0617	0.0588
8	35	0.0270	0.0250	19	56	0.0432	0.0400
9	56	0.0432	0.0400	20	35	0.0270	0.0250
10	80	0.0617	0.0588	21	20	0.0154	0.0143
11	104	0.0802	0.0794	22	10	0.0077	0.0075
12	125	0.0965	0.0984	23	4	0.0031	0.0036
13	140	0.1080	0.1119	24	1	0.0008	0.0016
14	146	0.1127	0.1168	Σ	1296	1.0000	0.9980

Aus dieser Häufigkeitsverteilung lässt sich der Durchschnitt $\mu := A(x) = 14$ und die Standardabweichung $\sigma := \sigma(x) = 3.42$ berechnen. In der obigen Tabelle haben wir bereits die tatsächlichen relativen Häufigkeiten $h(s)$ mit der Häufigkeitsdichte $g(\mu,\sigma,s)$ der Normalverteilung zu diesem Wertepaar (μ,σ) verglichen (die erst im folgenden gegeben wird). Man beachte, dass es sich bei der Abweichung *nicht* um Rundungsfehler, oder zufällige Abweichungen handelt. Die Abweichung ist systematisch - sie verschwindet erst, wenn die Anzahl m der Würfel unendlich $m \to \infty$ wird. Die Tatsache, dass die Gesamthäufigkeit der angenäherten Verteilung $g(\mu,\sigma,s)$ nicht 1 ergibt, liegt daran, dass es sich hierbei um eine Häufigkeits*dichte* handelt, die eben nicht zu einem diskreten Problem passt. Dass die Übereinstimmung aber trotzdem schon recht gut ist zeigt das folgende Histogramm, in dem die tatsächliche Häufigkeitsverteilung der 4 Würfel blau und die der Normalverteilung rot eingezeichnet ist.



Definition 13.1:
Seien $\mu, \sigma \in \mathbb{R}$ beliebige Parameter, dann definieren wir die **Gauss-Funktion** $g(\mu, \sigma) : \mathbb{R} \to \mathbb{R}^+$ zum *Mittelwert* μ und zur *Standardabweichung* σ durch

$$g(\mu, \sigma, x) \quad := \quad \frac{1}{\sigma \cdot \sqrt{2\pi}} \cdot \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Damit ist $g(\mu, \sigma)$ eine Dichtefunktion über $\mathbb{R}$ (d.h. das Integral von $g(\mu, \sigma)$ über $\mathbb{R}$ ist 1) und damit erhält man ein Wahrscheinlichkeitsmaß $\gamma(\mu, \sigma)$ auf der Borel-σ-Algebra $\mathcal{B}(\mathbb{R})$ durch (wobei $A \subseteq \mathbb{R}$ Borel-messbar)

$$\gamma(\mu, \sigma, A) \quad := \quad \int_A g(\mu, \sigma, t)dt$$

Die die zugehörige *Verteilungsfunktion* zu diesem Wahrscheinlichkeitsmaß wird mit $\Phi(\mu, \sigma) : \mathbb{R} \to [0, 1]$ bezeichnet. D.h. wir definieren

$$\Phi(\mu, \sigma, x) \quad := \quad \int_{-\infty}^{x} g(\mu, \sigma, t)dt$$

Insgesamt ist die **Normalverteilung** zum Mittelwert $\mu \in \mathbb{R}$ und zur Standardabweichung $\sigma > 0$ also der eben definierte Wahrscheinlichkeitsraum

$$\mathcal{N}(\mu, \sigma) \quad := \quad (\mathbb{R}, \mathcal{B}(\mathbb{R}), \gamma(\mu, \sigma))$$

Bemerkung 13.2:
Der Vorfaktor von $g(\mu, \sigma)$ ist so gewählt, dass das Integral von $g(\mu, \sigma)$ über ganz $\mathbb{R}$ gleich 1 wird (sonst wäre $\gamma(\mu, \sigma)$ kein Wahrscheinlichkeitsmaß)

$$\int_{\mathbb{R}} g(\mu, \sigma, t)dt \quad = \quad 1$$

Offensichtlich ist $g(\mu, \sigma)$ achsensymmetrisch, zu $x = \mu$, was dazu führt, das $\Phi(\mu, \sigma)$ punktsymmetrisch zu $(\mu \mid 1/2)$ ist und daher gilt für alle $x \in \mathbb{R}$

$$\Phi(\mu, \sigma, \mu + x) \quad = \quad 1 - \Phi(\mu, \sigma, \mu - x)$$

Des weiteren ergeben sich für $g(\mu, \sigma)$ die folgenden Ableitungen (nach der Variablen x, bei konstanten Parametern μ und σ)

$$\begin{aligned} \partial_x g(\mu, \sigma, x) &= -\frac{1}{\sigma^2}(x-\mu)g(\mu, \sigma, x) \\ \partial_x^2 g(\mu, \sigma, x) &= \frac{2}{\sigma^2}\left(\left(\frac{x-\mu}{\sigma}\right)^2 - 1\right) g(\mu, \sigma, x) \end{aligned}$$

Und damit kann man leicht nachrechnen, dass das Schaubild von $g(\mu, \sigma)$ den folgenden Hoch- bzw. die folgenden beiden Wendepunkte hat

$$\text{HP bei } \left(\mu \mid \frac{1}{\sigma^2\sqrt{2\pi}}\right) \quad \text{und} \quad \text{WP bei } \left(\mu \pm \sigma \mid \frac{1}{\sigma\sqrt{2e\pi}}\right)$$

Es bleibt die Frage, wie man $\Phi(\mu, \sigma)$ effektiv berechnen kann: dazu lässt sich die *Fehlerfunktion* erf heran ziehen, deren Reihenentwicklung man ausnutzen kann. Diese ist eng verwandt mit $g(0,1)$, denn es gilt

$$\begin{aligned} \operatorname{erf}(x) &:= \frac{2}{\sqrt{\pi}} \int_0^x \exp(-t^2) dt &= \sum_{k=0}^{\infty} \frac{(-1)^k}{(2k+1) \cdot k!} \cdot x^{2k+1} \\ \Longrightarrow \quad \Phi(\mu, \sigma, x) &= \frac{1}{2} \left(1 + \operatorname{erf}\left(\frac{x-\mu}{\sqrt{2}\sigma} \right) \right) \end{aligned}$$

Satz 13.3: Zentraler Grenzwertsatz
Sei (S, Σ, p) ein Wahrscheinlichkeitsraum, und $X_n \in \mathcal{Z}(S)$ (mit $n \in \mathbb{N}$) eine Folge von Zufallsvariablen auf S, so dass gilt

(1) alle X_n sind quadratisch p-integrierbar (d.h. es ist $\int X_n^2 \, dp < \infty$, näheres dazu findet sich im Kapitel 17 zur Integrationstheorie),

(2) die X_n sind identisch verteilt, d.h. für alle $m, n \in \mathbb{N}$ und alle $a < b \in \mathbb{R}$ gilt $p(X_m^{-1}([a,b[)) = p(X_n^{-1}([a,b[))$,

(3) die X_n sind unabhängig, d.h. für je $m \in \mathbb{N}$ endlich viele Zufallsvariablen $\{ i(1), \ldots, i(m) \} \subseteq \mathbb{N}$ und Borel-Mengen $B_k \in \mathcal{B}(\mathbb{R})$ gilt

$$p\left(\bigcap_{k=1}^{n} X_{i(k)}^{-1}(B_k) \right) = \prod_{k=1}^{n} p\left(X_{i(k)}^{-1}(B_k) \right)$$

Wir bezeichnen den (wegen (2) von n unabhängigen) Erwartungswert der X_n mit $\mu := E(X_n)$ und deren Varianz mit $\sigma^2 := V(X_n)$. Ferner setzten wir voraus, dass $\sigma \neq 0$ ist. Dann bezeichnen wir

$$S_n := X_1 + X_2 + \cdots + X_n \in \mathcal{Z}(S)$$

Damit ist S_n ebenfalls eine Zufallsvariable mit Erwartungswert $E(S_n) = n\mu$ und Varianz $V(S_n) = n\sigma^2$. Schließlich bezeichnen wir die normalisierte Zufallsvariable von S_n mit Z_n

$$Z_n := \frac{1}{\sigma\sqrt{n}} \cdot (S_n - n\mu) \in \mathcal{Z}(S)$$

d.h. es ist $E(Z_n) = 0$ und $V(Z_n) = 1$. Dann konvergiert die Verteilungsfunktion von Z_n punktweise gegen die Verteilungsfunktion $\Phi(0,1)$ der Standard-Normalverteilung. Explizit bedeutet das, dass für alle $x \in \mathbb{R}$ gilt:

$$\lim_{n \to \infty} p(Z_n \leq x) = \Phi(0, 1, x)$$

Noch expliziter ausgeschrieben bedeutet das, dass für zwei beliebige Zahlen $a \leq b \in \mathbb{R}$ stets die folgende Identität gilt

$$\lim_{n \to \infty} p\left(a \leq \frac{S_n - n\mu}{\sigma\sqrt{n}} \leq b \right) = \frac{1}{\sqrt{2\pi}} \int_a^b \exp\left(-\frac{t^2}{2} \right) dt$$

Satz 13.4: von Berry-Esseen
Wie oben sei $X_n \in \mathcal{Z}(S)$ wieder eine Folge ($n \in \mathbb{N}$) identisch verteilter, unabhängiger Zufallsvariabler auf dem Wahrscheinlichkeitsraum (S, Σ, p). Wir setzen wieder $\mu := E(X_n)$, $\sigma^2 := V(X_n)$ und fordern auch $0 < \sigma < \infty$. Ferner sei diesmal aber auch das zentrale dritte Moment τ der X_n endlich

$$\tau \;:=\; E\left(|X_n - \mu|^3\right) \;=\; \int_S |X_n(s) - \mu|^3 \, dp(s) \;<\; \infty$$

Dann lässt sich die Abweichung der Verteilung der Z_n (wie oben) von der Standard-Normalverteilung für alle $x \in \mathbb{R}$ und alle $n \in \mathbb{N}$ abschätzen, durch

$$|\, p(Z_n \leq x) - \Phi(0,1,x) \,| \;\leq\; \frac{3}{5} \cdot \frac{\tau}{\sigma^3 \sqrt{n}}$$

Bemerkung 13.5:
In der Anwendung ist es natürlich von Vorteil, wenn man die experimentellen Daten der Messungen $X_1, X_2, \ldots$ nicht erst normalisieren muss. Der zentrale Grenzwertsatz lässt sich aber auch leicht so transformieren, dass keine Normalisierung vorkommt: Sei wieder $S_n = X_1 + X_2 + \cdots + X_n$ und bezeichne $A_n := S_n/n$ das arithmetische Mittel der n ersten X_k. Dann erhalten für alle $x \in \mathbb{R}$ wir die Näherungen:

$$\begin{aligned} p(S_n \leq x) &\approx \Phi(n\mu, \sqrt{n}\sigma, x) \\ p(A_n \leq x) &\approx \Phi\left(\mu, \frac{\sigma}{\sqrt{n}}, x\right) \end{aligned}$$

Für den Beweis dieser transformierten Abschätzungen benötigen wir ein paar Verschiebungseigenschaften der Gauss-Kurve, respektive ihrer Verteilungsfunktion. Sei $\alpha > 0$ beliebig, dann gilt

$$g(\mu, \sigma, x) \;=\; \alpha \cdot g(\alpha\mu, \alpha\sigma, \alpha x) \;=\; \frac{1}{\sigma} \cdot g\left(0, 1, \frac{x-\mu}{\sigma}\right)$$

$$\Phi(\mu, \sigma, x) \;=\; \Phi(\alpha\mu, \alpha\sigma, \alpha x) \;=\; \Phi\left(0, 1, \frac{x-\mu}{\sigma}\right)$$

Beweis: Die erste Gleichung rechnet man einfach nach: in $g(\alpha\mu, \alpha\sigma, \alpha x)$ tritt der Term $(\alpha x - \alpha\mu)^2/2(\alpha\sigma)^2$ auf, kürzt man α^2 so bleibt nur noch ein Vorfaktor $1/\alpha$ übrig. Auch die zweite Gleichung rechnet man nach:

$$g\left(0, 1, \frac{x-\mu}{\sigma}\right) \;=\; \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\left(\frac{x-\mu}{\sigma}\right)^2}{2}\right) \;=\; \sigma \cdot g(\mu, \sigma, x)$$

Und damit folgt dann die Identität für Φ aus der Kettenregel, leiten wir ab:

$$\begin{aligned} \partial_x \Phi\left(0, 1, \frac{x-\mu}{\sigma}\right) &= \frac{1}{\sigma} (\partial_x \Phi)\left(0, 1, \frac{x-\mu}{\sigma}\right) \\ &= \frac{1}{\sigma} \cdot g\left(0, 1, \frac{x-\mu}{\sigma}\right) \;=\; g(\mu, \sigma, x) \end{aligned}$$

D.h. $\Phi(0,1,(x-\mu)/\sigma)$ ist eine Stammfunktion von $g(\mu,\sigma,x)$. Das gilt aber auch für $\Phi(\mu,\sigma,x)$, also unterscheiden sich beide nur um eine Konstante. Da sie aber auch im Grenzwert $x \to -\infty$ überein stimmen (beide konvergieren gegen 0), müssen sie also gleich sein. Mit demselben Argument zeigt man auch die zweite Identität für Φ: $\partial_x\Phi(\alpha\mu,\alpha\sigma,\alpha x) = \alpha(\partial_x\Phi)(\alpha\mu,\alpha\sigma,\alpha x) = \alpha \cdot g(\alpha\mu,\alpha\sigma,\alpha x) = g(\mu,\sigma,x)$. Es bleibt die Näherungen zu beweisen, diese kann man nunmehr aber ebenfalls nachrechnen:

$$\begin{aligned}
p(S_n \le x) &= p\{\, s \in S \mid S_n(s) \le x \,\} \\
&= p\left\{ s \in S \;\middle|\; \frac{S_n(s)-n\mu}{\sigma\sqrt{n}} \le \frac{x-n\mu}{\sigma\sqrt{n}} \right\} \\
&= p\left(Z_n \le \frac{x-n\mu}{\sigma\sqrt{n}} \right) \;\approx\; \Phi\left(0,1,\frac{x-n\mu}{\sigma\sqrt{n}}\right) \\
&= \Phi(n\mu,\sqrt{n}\sigma,x)
\end{aligned}$$

Zu guter Letzt betrachten wir $p(A_n \le x) = p(S_n/n \le x) = p(S_n \le nx) \approx \Phi(n\mu,\sqrt{n}\sigma,nx)$. Wählen wir $\alpha = n$ dann lässt sich dies wie behauptet schreiben, als $p(A_n \le x) \approx \Phi(\alpha\mu,\alpha(\sigma/\sqrt{n}),\alpha x) = \Phi(\mu,\sigma/\sqrt{n},x)$.

□

Beispiel 13.6:

(i) Das klassische Beispiel ist die vielfache Wiederholung immer ein und desselben Zufallsexperimentes. Wir starten also mit einem fixiertem Wahrscheinlichkeitsraum (S,Σ,p) und einer Zufallsvariablen $X \in \mathcal{Z}(S)$. Wir setzten voraus, dass X quadratisch p-integrierbar ist und setzen $\mu := E(X)$ und $\sigma^2 := V(X)$. Die vielfache Wiederholung von X wird modelliert durch den $\mathbb{N}$-fachen Produktraum

$$(S,\Sigma,p)^{\otimes\mathbb{N}}$$

D.h. wir betrachten Folgen $(s_i) \in S^{\mathbb{N}}$ mit Einträgen aus S, die die Ergebnisse einer unendlich wiederholten Ziehung aus S (mit Zurücklegen) darstellen. Sei X_k die Zufallsvariable X in der k-ten Ziehung

$$X_k \;:\; S^{\mathbb{N}} \to \mathbb{R} \;:\; (s_i) \mapsto X(s_k)$$

dann sind die X_k wieder (wie immer in diesem Fall) identisch verteilt (sie haben dieselbe Verteilung wie X selbst) und sind unabhängig voneinander. Die S_n sind dann gegeben, durch

$$S_n \;:\; S^{\mathbb{N}} \to \mathbb{R} \;:\; (s_i) \mapsto X(s_1) + \cdots + X(s_n)$$

so dass hier auch $E(S_n) = n\mu$ und $V(S_n) = n\sigma^2$ besonders einsichtig wird. Nun ist aber die Voraussetzung $\sigma \neq 0$ wichtig! X muss eine Streuung der Ergebnisse haben, kommt (fast sicher) immer dasselbe Ergebnis heraus, so können sich die Ergebnisse nicht gegenseitig aufheben.

Ist dies aber gegeben, dann besagt der zentrale Grenzwertsatz bereits, dass die Verteilung der Z_n punktweise gegen $\Phi(0,1)$ konvergiert. Das heißt für hinreichend große n kann man schätzen

$$p\left(a \leq \frac{S_n - n\mu}{\sigma\sqrt{n}} \leq b\right) \approx \frac{1}{\sqrt{2\pi}} \int_a^b \exp\left(-\frac{t^2}{2}\right) dt$$

Bezeichnen wir wieder das arithmetische Mittel $A_n := S_n/n$, dann lautet die rücktransformierte Abschätzung, wie eben gezeigt wurde:

$$p(a \leq A_n \leq b) \approx \frac{1}{\sigma} \cdot \sqrt{\frac{n}{2\pi}} \int_a^b \exp\left(-n\frac{(t-\mu)^2}{2\sigma^2}\right) dt$$

(ii) Ein Spezialfall des generellen Problems (i) ist die Binomialverteilung: wir betrachten den n-maligen Münzwurf $\mathcal{B}(n,\lambda)$ wobei $0 < \lambda < 1$ sein muss. Mit $\omega : \{0,1\}^n \to 0 \ldots n$ bezeichnen wir wieder die Zufallsvariable der Zählung der 1 in der Folge $\omega(s) = \#\{i \in 1 \ldots n \mid s_i = 1\}$. Wie wir wissen ist dann $\mu = E(\omega) = n\lambda$ und $\sigma^2 = V(\omega) = n\lambda(1-\lambda)$. Man beachte, dass μ und σ^2 hier bereits Erwartungswert und Varianz des n-fachen Münzwurfs (und damit das n-fache des einfachen Münzwurfs) sind. Es hat sich etabliert zu sagen, dass sich die Binomialverteilung mit genügender Genauigkeit durch die Normalverteilung approximieren lässt, wenn $\sigma^2 > 9$ ist. Dann gilt für $k \in 0 \ldots n$ die Näherung

$$p(\omega = k) \approx g(\mu,\sigma,k) = \frac{1}{\sqrt{2\pi n\lambda(1-\lambda)}} \cdot \exp\left(-\frac{(k-n\lambda)^2}{2n\lambda(1-\lambda)}\right)$$

Will man einen ganzen Abschnitt $0 \leq a \leq b \leq n$ möglicher Anzahlen von 1en auswerten, so ist die Approximation besser, wenn man eine sogenannte *Stetigkeitskorrektur* durchführt, soll heißen das betrachtete Intervall um jeweils 0.5 auf beiden Seiten erweitert

$$\begin{aligned} p(\omega \leq b) &\approx \Phi(\mu,\sigma,b+0.5) \\ p(a \leq \omega \leq b) &\approx \Phi(\mu,\sigma,b+0.5) - \Phi(\mu,\sigma,a-0.5) \end{aligned}$$

Die obige Näherungsformel für $p(\omega = k)$ wird auch Satz von *Moivre-Laplace* genannt, nach dem das Verhältnis der beiden Terme im Grenzwert $n \to \infty$ gegen 1 geht. In den Beweis dieses Satzes gehen die Stirling'sche Formel mehrfach und auch die Reihenentwicklung von $\ln(x+1)$ ein. Da $k = \mu + x\sigma$ betrachtet wird, ist die Approximation umso besser, je näher k an μ liegt.

Allgemeiner kann man den Satz von Berry-Esseen verwenden, für den man aber das zentrale, absolute dritte Moment der Binomialverteilung benötigt: dieses ist gegeben, durch $\tau(\lambda) = \sigma^2(\lambda) \cdot (1 - a\lambda + a\lambda^2)$ wobei $\sigma^2(\lambda) = n\lambda(1-\lambda)$ wieder die Varianz und $a = 4 - 16\tau(0.5)/n$ ist. Da sowohl $\sigma^2(\lambda)$ also auch $1 - a\lambda + a\lambda^2$ ihr Maximum bei $\lambda = 1/2$ annehmen, gilt stets $\tau(\lambda) \leq \tau(0.5)$ oder genauer

$$\tau(\lambda) = \sigma^2(\lambda) \cdot (1 - a\lambda + a\lambda^2) \leq \frac{n}{4}\left(1 - \frac{a}{4}\right) = \tau(0.5)$$

Um diese Abschätzung anwenden zu können benötigt man also noch das zentrale, absolute dritte Moment $\tau(0.5)$ der Binomialverteilung für $\lambda = 0.5$. Und dieses lässt sich für $n = 2m$ (obere Formel) bzw. für $n = 2m + 1$ (untere Formel) berechnen, durch:

$$\tau(0.5) \;=\; \frac{1}{2^n}(m+1)m\binom{n}{m+1} \;<\; \frac{1}{5}n^{1.5}$$

$$\tau(0.5) \;=\; \frac{1}{2^n}(m+1)\left(m+\frac{1}{4}\right)\binom{n}{m+1} \;<\; \frac{1}{5}n^{1.5}$$

In der Schranke von Berry-Esseen kann man $\tau(\lambda) \leq \tau(0.5) < n^{1.5}/5$ abschätzen. Alternativ kann man auch $\tau(\lambda)/\sigma(\lambda)^2 = 1 - a\lambda + a\lambda^2 \leq 1 - a/4 = 4\tau(0.5)/n < (4/5) \cdot \sqrt{n}$ abschätzen und erhält so schließlich

$$|\, p(Z_n \leq x) - \Phi(0,1,x)\,| \;\leq\; \frac{3}{5} \cdot \frac{\tau(0.5)}{\sigma(\lambda)^3 \cdot \sqrt{n}} \;<\; \frac{12}{25 \cdot \sigma(\lambda)}$$

(iii) Für kleine Parameterwerte λ ist die Poissonverteilung $\mathcal{P}(\lambda) = (\mathbb{N}, p_\lambda)$ stark asymmetrisch, für größere Werte jedoch - sagen wir ab $\lambda \geq 30$ - ähnelt die Poisson-Verteilung zunehmend einer Normalverteilung vom Mittelwert $\mu = \lambda$ und mit der Varianz $\sigma^2 = \lambda$. In diesem Fall können wir also die folgenden Approximationen (man beachte wieder die Stetigkeitskorrektur) verwenden (wobei $k \in \mathbb{N}$ beliebig ist):

$$\begin{aligned} p_\lambda(\mathrm{id} = k) &\approx g(\lambda, \sqrt{\lambda}, k) \\ p_\lambda(\mathrm{id} \leq k) &\approx \Phi(\lambda, \sqrt{\lambda}, k + 0.5) \end{aligned}$$

Da für die Poisson-Verteilung auch das dritte zentrale Moment existiert (es ist auch $\tau = \lambda$) kann man den Satz von Berry-Esseen wieder zur Abschätzung des Fehlers heran ziehen. Dabei wird ebenfalls deutlich (für $n = 1$), dass sich $\mathcal{P}(\mathbb{N}, p_\lambda)$, mit wachsendem λ, $\mathcal{N}(\lambda, \sqrt{\lambda})$ nähert:

$$|\, p_\lambda(Z_n \leq k) - \Phi(0,1,k)\,| \;\leq\; \frac{\lambda}{2\lambda^{3/2}\sqrt{n}} \;=\; \frac{1}{2 \cdot \sqrt{\lambda n}}$$

Bemerkung 13.7:
Sei (X, Σ, p) ein Wahrscheinlichkeitsraum, dann nennen wir eine Zufallsvariable $X \in \mathcal{Z}(S)$ **normalverteilt** falls es $\mu \in \mathbb{R}$ und $\sigma > 0$ gibt, so dass die Verteilungsfunktion von X gleich der Verteilungsfunkton von $\gamma(\mu, \sigma)$ ist

$$F_X \;=\; \Phi(\mu, \sigma)$$

Sind nun X und $X' \in \mathcal{Z}(S)$ zwei unabhängige, normalverteilte Zufallsvariablen mit $F_X = \Phi(\mu, \sigma)$ und $F_{X'} = \Phi(\mu', \sigma')$ dann ist auch die Summe $X + Y$ der beiden normalverteilt [für einen Beweis, siehe www.vwi.tu-dresden.de/~treiber/statistik2/statistik_download/exkurse15.pdf] mit

$$F_{X+Y} \;=\; \Phi\left(\mu + \mu', \sqrt{\sigma^2 + \sigma'^2}\right)$$

χ^2-Verteilung:
Wie wir gerade gesehen haben, ist die Summe $X_1 + \cdots + X_n$ normalverteilter Zufallsvariablen X_1 bis $X_n \in \mathcal{Z}(S)$ wieder normalverteilt. Für Hypothesentests benötigt man allerdings nicht die Summe, sondern die Summe der Quadrate $X_1^2 + \cdots + X_n^2$. Wir nehmen zunächst an, alle X_i seien standardnormalverteilt, d.h. es sei stets $X_i \sim \mathcal{N}(0,1)$. Die Verteilung zu $X_1^2 + \cdots + X_n^2$ hängt dann nur von der Zahl n der beteiligten Zufallsvariablen ab. Wir wollen diese sogenannte χ^2-Verteilung nun im folgenden studieren:

Bemerkung 13.8:
Zunächst benötigen wir die **Gamma-Funktion**, welche eine Verallgemeinerung der Fakultät darstellt: Wir definieren $\Gamma : 0 \to \mathbb{R}^+$ durch:

$$\Gamma(x) \;:=\; \int_0^\infty t^{x-1} e^{-t} dt$$

Dann lässt sich zeigen, dass Γ eine (nicht eindeutig bestimmte) Lösung der folgenden Funktionalgleichung (für alle $x \geq 0$) ist:

$$\Gamma(1) \;=\; 1 \quad \text{und} \quad \Gamma(x+1) = x \cdot \Gamma(x)$$

Insbesondere folgt daraus $\Gamma(n+1) = n!$ für alle $n \in \mathbb{N}$. Da wir Γ im folgenden aber nur auf halbzahligen Werten berechnen müssen, genügen uns eigentlich bereits die expliziten Formeln (mit $1 \leq n \in \mathbb{N}$):

$$\begin{aligned} \Gamma(n) &= (n-1)! \\ \Gamma(0.5) &= \sqrt{\pi} \\ \Gamma(n+0.5) &= \frac{\sqrt{\pi}}{2^n} \prod_{k=1}^{n} (2k-1) \end{aligned}$$

Definition 13.9:
Sei nun $1 \geq n \in \mathbb{N}$ vorgelegt, dann definieren wir die folgende Dichtefunktion

$$c_n \;:\; \mathbb{R}^+ \to \mathbb{R} \;:\; x \;\mapsto\; \frac{x^{\frac{n}{2}-1} e^{-\frac{x}{2}}}{2^{\frac{n}{2}} \Gamma\left(\frac{n}{2}\right)}$$

Oftmals wird c_n (stetig) erweitert, auf $c_n : \mathbb{R} \to \mathbb{R}$ indem man für $x \leq 0$ definiert $c_n(x) := 0$. Damit erhält man dann ein Wahrscheinlichkeitsmaß auf $\mathbb{R}^+$ durch (wobei $A \subseteq \mathbb{R}^+$ Borel-meßbar)

$$\chi_n(A) \;:=\; \int_A c_n(t) dt$$

Die Verteilungsfunktion dieses Wahrscheinlichkeitsmaßes bezeichnen wir (etwas informal, aber sie ist ja auch fast dasselbe) ebenfalls mit (wobei $x \geq 0$)

$$\chi_n(x) \;:=\; \int_0^x c_n(t) dt$$

Der zugehörige Wahrscheinlichkeitsraum wird als **Raum der χ^2-Verteilung** mit n **Freiheitsgraden** bezeichnet, wir schreiben diesen, als

$$\mathcal{C}(n) \quad := \quad (\mathbb{R}^+, \mathcal{B}(\mathbb{R}) \cap \mathbb{R}^+, \chi_n)$$

Satz 13.10:

(i) Bezeichnen wir die Zufallsvariable $\mathrm{id} : \mathbb{R}^+ \to \mathbb{R}^+ : x \mapsto x$ der Identität, dann hat diese den folgenden Erwartungswert, bzw. die folgende Varianz über der χ^2-Verteilung $\mathcal{C}(n)$

$$\begin{aligned} E(\mathrm{id}) &= n \\ V(\mathrm{id}) &= 2n \end{aligned}$$

Die erzeugende Funktion der Momente lässt sich ebenfalls berechnen:

$$M_{\mathrm{id}}(t) \quad = \quad (1-2t)^{-\frac{n}{2}}$$

(ii) Für $n = 1$ lässt sich die Verteilungsfunktion χ_n mit Hilfe der (im Anschluss an die Normalverteilung eingeführten) erf-Funktion berechnen:

$$\chi_1(x) \;=\; \mathrm{erf}\left(\sqrt{\frac{x}{2}}\right) \;=\; \sum_{k=0}^{\infty} \frac{(-1)^k}{(2k+1)\cdot k!} \cdot \left(\frac{x}{2}\right)^{k+0.5}$$

Für gerade n mit $n \geq 2$ setzen wir $m := n/2 - 1$, dann ergibt sich die folgende Formel für die Verteilungsfunktion der χ^2-Verteilung:

$$\chi_n(x) \quad = \quad 1 - e^{-\frac{x}{2}} \sum_{k=0}^{m} \frac{1}{\Gamma(k+1)} \left(\frac{x}{2}\right)^k$$

Und für ungerade n mit $n \geq 3$ setzen wir $m := (n-3)/2$, dann ergibt sich eine etwas andere Formel für die Verteilungsfunktion

$$\chi_n(x) \quad = \quad \mathrm{erf}\left(\sqrt{\frac{x}{2}}\right) - e^{-\frac{x}{2}} \sum_{k=0}^{m} \frac{1}{\Gamma(k+1.5)} \left(\frac{x}{2}\right)^{k+0.5}$$

(iii) Sei (S, Σ, p) ein beliebiger Wahrscheinlichkeitsraum und seien $Z(1)$, $Z(2)$ bis $Z(n) \in \mathcal{Z}(S)$ unabhängige Zufallsvariablen auf S. Ferner seien alle $Z(i)$ standard-normal-verteilt, d.h. für alle $i \in 1 \ldots n$ gelte $F_{Z(i)} = \Phi(0,1)$. Dann ist $Q(n) := Z(1)^2 + Z(2)^2 + \cdots + Z(n)^2 \in \mathcal{Z}(S)$ (also die Summe der Quadrate der $Z(i)$) gerade χ^2-verteilt, d.h. es gilt

$$F_{Q(n)} \quad = \quad \chi_n$$

(iv) Sind $1 \leq m$, $n \in \mathbb{N}$, dann sind die unabhängigen Kombinationen der χ^2-Verteilungen $\mathcal{C}(m)$ und $\mathcal{C}(n)$ (aufgrund von (iii) offenbar) wieder eine χ^2-Verteilung, nämlich $\mathcal{C}(m+n)$

$$\mathcal{C}(m) \otimes \mathcal{C}(n) \quad = \quad \mathcal{C}(m+n)$$

Bemerkung 13.11:
Es seien $X(1)$, $X(2), \ldots, X(n) \in \mathcal{Z}(S)$ unabhängige, normal-verteilte Zufallsvariablen auf dem Wahrscheinlichkeitsraum (S, Σ, p) mit den Erwartungswerten $\mu(i) = E(X(i)) \in \mathbb{R}$ und Varianzen $\sigma(i)^2 = V(X(i)) \neq 0$, d.h. es sei jeweils

$$F_{X(i)} \quad = \quad \Phi(\mu(i), \sigma(i))$$

Dann betrachten wir zuerst die Zufallsvariable, die den Durchschnitt der $X(i)$ beschreibt. Diese ist offenbar gegeben, durch:

$$A(n) \quad := \quad \frac{1}{n} \sum_{i=1}^{n} X(i)$$

Aufgrund der Linearität des Erwartungswertes und der Additivität der Varianz (auf zumindest paarweise unabhängignen Zufallsvariablen) können wir eben diese Größen auch für $A(n)$ berechnen:

$$\begin{aligned} E(A(n)) \quad &= \quad \frac{1}{n} \sum_{i=1}^{n} E(X(i)) \quad = \quad \frac{1}{n} \sum_{i=1}^{n} \mu(i) \\ V(A(n)) \quad &= \quad \frac{1}{n^2} \sum_{i=1}^{n} V(X(i)) \quad = \quad \frac{1}{n^2} \sum_{i=1}^{n} \sigma(i)^2 \end{aligned}$$

Insgesamt ist $A(n)$ damit wieder normal-verteilt, mit Mittelwert $\mu = E(A(n))$ und Standardabweichung $\sigma = \sqrt{V(A(n))}$. Wir wollen die $X(i)$ nun aber *standardisierten*, d.h. so transformieren, dass die neue Zufallsvariable $Z(i)$ standard-normalverteilt ist. Dies geschieht unter:

$$Z(i) \quad := \quad \frac{X(i) - \mu(i)}{\sigma(i)}$$

Damit sind die $Z(i)$ weiterhin unabhängig und nunmehr auch standard-normal-verteilt, $F_{Z(i)} = \Phi(0, 1)$ und es besteht der Zusammenhang

$$p(X(i) \leq x) \quad = \quad p\left(Z(i) \leq \frac{x - \mu(i)}{\sigma(i)}\right) \quad = \quad \Phi\left(0, 1, \frac{x - \mu(i)}{\sigma(i)}\right)$$

$$p(a < Z(i) < b) \quad = \quad p\big(a\sigma(i) + \mu(i) < X(i) < a\sigma(i) + \mu(i)\big)$$

für beliebige a, b und $x \in \mathbb{R}$. Aufgrund des obigen Satzes wissen wir nun aber auch, dass $Q(n) := Z(1)^2 + \cdots + Z(n)^2$ χ^2-verteilt ist: d.h. es gilt $F_{Q(n)} = \chi_n$, oder explizit gilt für alle $x \geq 0$:

$$p\left(\left\{ s \in S \;\middle|\; \sum_{i=1}^{n} Z(i)(s)^2 \leq x \right\}\right) \quad = \quad \chi_n(x) \quad = \quad \int_0^x c_n(t)dt$$

Satz 13.12: Pearson's Theorem:
Sei (S, Σ, p) ein Wahrscheinlichkeitsraum und (T, q) ein Laplace-Raum, mit $T = \{ t_1, \ldots, t_m \}$. Sei ferner (X_k) eine Folge unabhängiger Zufallsvariablen $X_k : S \to T$, so dass für alle $k \in \mathbb{N}$ und alle $j \in 1 \ldots m$ gilt:

$$p\big(X_k = t_j\big) \;=\; q_j$$

Zu $s \in S$ definieren wir $N_{j,n}(s) := \#\{ k \in 1 \ldots n \mid X_k(s) = t_j \}$. Dann ist $N_{j,n} : S \to \mathbb{R}$ eine reelle Zufallsvariable und wir definieren eine weitere Zufallsvariable $\Delta_n : S \to \mathbb{R}$ durch

$$\Delta_n(s) \;:=\; \sum_{j=1}^{m} \frac{\big(N_{j,n}(s) - nq_j\big)^2}{nq_j}$$

Dann konvergiert (Δ_n) in Verteilung gegen die Chi-Quadrat-Verteilung mit $m-1$ Freiheitsgraden. D.h. ist $F_n(x) = p(\Delta_n \leq x)$ die Verteilungsfunktion von Δ_n dann konvergiert diese punktweise gegen χ_{m-1}: Für alle $x \in \mathbb{R}$ gilt:

$$\lim_{n\to\infty} F_n(x) \;=\; \chi_{m-1}(x)$$

Äquivalent dazu: Für alle $a < b \in \mathbb{R}$ gilt

$$\lim_{n\to\infty} p\big(a \leq \Delta_n \leq b\big) \;=\; \chi_{m-1}(b) - \chi_{m-1}(a)$$

Ein Freiheitsgrad verschwindet, denn kennt man die Häufigkeit des Vorkommens von t_2 bis t_m, dann ist auch die Häufigkeit von t_1 klar. Einen Beweis dieses Satzes findet man z.B. in den Open Courseware Notes des MIT zur Statistik, Section 10. Der Nutzen des Satzes besteht darin, dass die Abweichung der Verteilungsfunktion F_n von χ_{m-1} zu einem Test umwandeln, ob die Wahrscheinlichkeiten $p(X_k = t_j)$ wirklich q_j betragen: Ist $x = (x_1, \ldots, n_x) \in T^n$ eine Stichprobe, dann bezeichnen wir mit $n_j := \#\{ k \in 1 \ldots n \mid x_k = t_j \}$ die absolute Häufigkeit von t_j in x. Damit berechnen wir dann

$$\Delta \;:=\; \sum_{j=1}^{m} \frac{\big(n_j - nq_j\big)^2}{nq_j}$$

Für $n \to \infty$ wird die relative Häufigkeit $h_j = n_j/n$ gegen die tatsächliche Wahrscheinlichkeit $p(X = t_j)$ konvergieren. Also sollte $n_j \approx nq_j$ sein. Und sind die q_j die richtigen Wahrscheinlichkeiten, dann kontrolliert Δ die empirische Verteilungsfunktion, die punktweise gegen χ_{m-1} konvergiert. Wenn die Abweichung zu groß ist, dann sind die q_j also die falschen Wahrscheinlichkeiten. Die genaue Ausführung des Tests wird in Kapitel 12 besprochen. Hat man hingegen keine Vermutung über die genaue Verteilung q auf T, sondern nur über ein Modell (T, q_λ) und möchte den Parameter λ schätzen, so hilft der folgende Satz:

Satz 13.13:
Sei wieder (S, Σ, p) ein Wahrscheinlichkeitsraum, $T = \{ t_1, \ldots, t_m \}$ eine endliche Menge und $\Lambda \subseteq \mathbb{R}^r$. Zu jedem $\lambda \in \Lambda$ sei (T, q_λ) ein Laplace-Raum (also $q_\lambda \in \Delta_m$). Und zu gegebenem $j \in 1 \ldots m$ betrachten wir

$$q_j \ : \ \Lambda \to [0,1] \ : \ \lambda \mapsto q_\lambda(t_j)$$

Sei weiterhin $X : S \to T$ eine Zufallsvariable und $(x_k) \subseteq T$ eine Folge von Ziehungen von X. Bezeichne $h_n \in \Delta_m$ den MLE-Schätzer der Häufigkeitsverteilung, der sich aus der Stichprobe $(x_1, \ldots, x_n)$ ergibt, d.h.

$$\begin{aligned} h_{j,n} &:= \frac{1}{n} \# \{ k \in 1 \ldots n \mid x_k = t_j \} \\ h_n &:= \big(h_{1,n}, \ldots, h_{m,n} \big) \\ h &:= \lim_{n \to \infty} h_n \end{aligned}$$

Damit konvergiert (h_n) gegen h, die tatsächliche Wahrscheinlichkeitsverteilung von X auf T. Sei $\vartheta \in \Lambda$ so dass $q_\vartheta = h$. Wir setzen voraus, dass

(1) ϑ ist ein innerer Punkt von Λ

(2) Für alle $j \in 1 \ldots m$ ist $q_j(\vartheta) \neq 0$.

(3) Für alle $j \in 1 \ldots m$ ist $q_j : \Lambda \to [0,1]$ C^2-differenzierbar.

(4) Die $(r \times m)$-Jacobi-Matrix $(\partial_{\lambda_i} q_j)$ hat bei $\lambda = \vartheta$ den vollen Rang r.

Wie in (13.12) definieren wir $N_{j,n}(s) := \# \{ k \in 1 \ldots n \mid X(s_k) = t_j \}$ für $s = (s_k) \in S^{\mathbb{N}}$ und damit die Zufallsvariable Δ_n auf $(S, \Sigma, p)^{\otimes \mathbb{N}}$

$$\Delta_n(s) \ := \ \sum_{j=1}^{m} \frac{\big(N_{j,n}(s) - n h_{j,n} \big)^2}{n h_{j,n}}$$

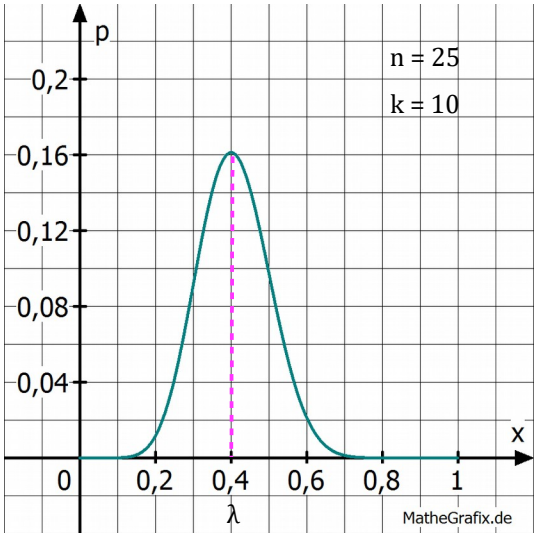
Dann konvergiert die Verteilungsfunktion F_n von Δ_n wieder punktweise gegen die Chi-Quadrat-Verteilung zu $f = m - r - 1$ Freiheitsgraden:

$$\lim_{n \to \infty} p^{\otimes \mathbb{N}} \big(\Delta_n \leq t \big) \ = \ \chi_f(t)$$

Kapitel 14

Konfidenzintervalle

Im folgenden möchten wir uns einem Problem zuwenden, das nur auf den ersten Blick einfach aussieht: Nehmen wir an, in einer Urne befinden sich insgesamt 100 Kugeln, der Farben schwarz und weiß. Wir ziehen $n = 25$ mal mit Zurücklegen (d.h. ziehen, Farbe notieren, Kugel zurücklegen und mischen). Dabei sind $k = 10$ der gezogenen Kugeln schwarz. Wie viele schwarze Kugeln werden wohl in der Urne sein? Naiv würde man rechnen: wir haben $\lambda = k/n = 0.4 = 40\%$ schwarze Kugeln gezogen, also sind wohl $100 \cdot \lambda = 40$ schwarze Kugeln in der Urne. Wir wissen bereits, dass dieses λ der Maximum Likelihood-Schätzer des Problems ist - d.h. unter allen relativen Häufigkeiten $\lambda \in [0, 1]$ hat $\lambda = 0.4$ die höchste Wahrscheinlichkeit das Ergebnis 10 schwarze Kugeln zu liefern:



In diesem Schaubild ist x die tatsächliche (unbekannte) realtive Häufigkeit schwarzer Kugeln in der Urne und auf der p-Achse haben wir die Wahrscheinlichkeit für $k = 10$ schwarzer unter $n = 25$ gezogener Kugeln aufgetragen. Wenn wir also erfahren, dass nur 35 Kugeln schwarz sind, dann überrascht uns das wenig - wir haben ja zufällig gezogen und auch dies liefert häufig genug 10 schwarze Kugeln. Wenn uns aber gesagt wird, das es seien nur 5 Schwarze, dann wäre das doch eine sehr seltene Ziehung gewesen. Oder gar 99 weiße Kugeln? Haben wir 10 Mal genau diese eine schwarze Kugel erwischt?

Letzteres ist zwar theoretisch möglich, aber sehr sehr fragwürdig. Doch das einzige, was wir wirklich sicher wissen, ist, dass die Zahl schwarzer Kugeln zwischen 1 und 99 liegen muss. Also: wie viele schwarze Kugeln sollen wir noch als glaubwürdig zulassen und ab wie vielen Kugeln sollen wir die Zahl als zu unwahrscheinlich verwerfen?

Formulieren wir das Problem etwas präziser: vorgelegt sei eine erlaubte Irrtumswahrscheinlichkeit $\alpha \in [0,1]$. In welchem Intervall $C_0(n,k,\alpha) \subseteq [0,1]$ liegt die tatsächliche Häufigkeit x der schwarzen Kugeln mit einer Sicherheit von $1-\alpha$. Dazu müssen wir den Begriff *Sicherheit* noch etwas durchdenken: Ein Irrtum liegt dann vor, wenn $x \notin C_0(n,k,\alpha)$ ist. Führen wir den Versuch ($n = 25$ Kugeln mit Zurücklegen ziehen, die Zahl k der gezogenen schwarzen Kugeln notieren und das *Konfidenzintervall* $C_0(n,k,\alpha)$ berechnen) immer wieder durch, dann kommt das jeweilige $k \in 0 \ldots n$ mit der Häufigkeit vor

$$p_k(x) \;:=\; b(n,x,k) \;=\; \binom{n}{k} x^k (1-x)^{n-k}$$

Ist $x \in C_0(n,k,\alpha)$ für dieses k erfüllt, dann hat unser Konfidenzintervall richtig gelegen und es ergibt sich kein Beitrag zur Irrtumswahrscheinlichkeit. Wir müssen also p_k über all die k summieren, für die $x \notin C_0(n,k,\alpha)$ ist - die Irrtumswahrscheinlichkeit kann also berechnet werden, zu

$$i(x,n) \;=\; \sum_{k=0}^{n} \begin{cases} p_k(x) & \text{für } x \notin C_0(n,k,\alpha) \\ 0 & \text{für } x \in C_0(n,k,\alpha) \end{cases}$$

Unsere Bedingung, die die Konfidenzintervalle $C_0(n,k,\alpha)$ erfüllen müssen ist also $i(x,n) \leq \alpha$. Eine Gleichheit ist nicht zu erwarten, da $i(x,n)$ ja nur in diskreten Schritten vorkommt. Die Antwort $C_0(n,k,\alpha) = [0,1]$ würde die Bedingung natürlich erfüllen, ist aber auch nicht wirklich hilfreich. Wir wollen die Konfidenzintervalle $C_0(n,k,\alpha)$ so konstruieren, dass $i(x,n) = \alpha$ tatsächlich gerade noch vorkommen kann.

Eine richtig befriedigende Antwort auf diese Frage gibt es leider nicht. Man kann ein exaktes Intervall angeben, das aber nur näherungsweise zu berechnen ist und man kann exakt zu berechnende Intervalle angeben, die aber nur näherungsweise treffen. Einen Tod müssen wir sterben und tasten uns daher nur langsam an eine Lösung heran. Zuletzt geben wir noch eine allgemeinere Formulierung des Problems an, jedoch ohne einen echten Lösungsalgorithmus zu präsentieren.

1. Versuch: Umkehrung der Wahrscheinlichkeitsfunktion

Betrachtet man die Irrtumswahrscheinlichkeit $i(x,n)$ ist klar, dass wir die Wahrscheinlichkeiten $p_k(x) = p_x(\omega = k)$ für das Eintreffen von k Erfolgen bei n Versuchen unter der unbekannten Erfolgswahrscheinlichkeit x kontrollieren müssen - siehe auch Kapitel 8 für die Frage, warum dies die richtige Wahrscheinlichkeit ist. Die Idee ist krude, aber effektiv: wir wissen nicht wo x liegt, aber wenn wir stets die Größe von $p_k(x)$ kontrollieren, dann kontrollieren wir auch die Irrtumswahrscheinlichkeit $i(x,n)$. Dies leisten nun die folgenden beiden Sätze. Der erste ist eine Untersuchung der Eigenschaften von $p_k(x)$ und mündet in den zweiten Satz - einem ersten Verfahren zur Berechnung von Konfidenzintervallen.

Satz 14.1:
Sei $1 \leq n \in \mathbb{N}$ und $k \in 0 \ldots n$ fixiert und bezeichne $\lambda := k/n$. Dann betrachten wir das folgende Polynom n-ten Grades mit ganzzahligen Koeffizienten

$$p_k(x) \quad := \quad \binom{n}{k} x^k (1-x)^{n-k} \quad \in \quad \mathbb{Z}[x]$$

- Für $k = 0$ hat p_k die Nullstelle $x = 1$ und ist auf $[0, 1]$ streng monoton fallend von $(0, 1)$ nach $(0, 0)$.
- Für $k = n$ hat p_k die Nullstelle $x = 0$ und ist auf $[0, 1]$ streng monoton steigend von $(0, 0)$ nach $(1, 1)$.
- Für $0 < k < n$ hat p_k die Nullstellen $x = 0$ und $x = 1$, ist über $[0, \lambda]$ streng monoton steigend, erreicht bei $x = \lambda$ einen Hochpunkt und ist über $[\lambda, 1]$ streng monoton fallend. Dabei ist

$$p_k(\lambda) \quad = \quad \binom{n}{k} \frac{k^k (n-k)^{n-k}}{n^n}$$

- Ist $k \geq 1$ und bezeichnet $q_k(x)$ das nachfolgende Polynom für $n + 1$, dann gelten die folgenden beiden Rekursionsformeln:

$$\begin{aligned} q_k(x) \quad &:= \quad \binom{n+1}{k} x^k (1-x)^{n+1-k} \\ p_{k-1}(x) \quad &= \quad \frac{k}{n-k+1} \cdot \frac{1-x}{x} \cdot p_k(x) \\ q_k(x) \quad &= \quad \frac{n+1}{k} \cdot x \cdot p_{k_1}(x) \end{aligned}$$

  Insbesondere schneidet p_{k-1} den Nachfolger q_k in dessen Hochpunkt:

$$q_k\left(\frac{k}{n+1}\right) \quad = \quad p_{k-1}\left(\frac{k}{n+1}\right)$$

Beweis:
Im Fall $k = 0$ ist $p_0(x) = (1-x)^n$ und erfüllt damit offensichtlich die oben genannten Eigenschaften. Desgleichen für $k = n$, hierbei ist $p_n(x) = x^n$. Sei also $1 < k < n$ und $\lambda = k/n$. Die Ableitung von $p_k(x)$ ergibt sich nach der Produktregel zu

$$\begin{aligned} p_k'(x) \quad &= \quad \binom{n}{k} k x^{k-1} (1-x)^{n-k} - \binom{n}{k} (n-k) x^k (1-x)^{n-k-1} \\ &= \quad \binom{n}{k} x^{k-1} (1-x)^{n-k-1} (k(1-x) - (n-k)x) \\ &= \quad \binom{n}{k} x^{k-1} (1-x)^{n-k-1} (k - nx) \end{aligned}$$

Auf $[0,1]$ ist sowohl x^{k-1} also auch $(1-x)^{n-k-1}$ stets positiv, das Vorzeichen von $p_k'(x)$ wird also durch den Term $(k-nx)$ bestimmt. Und für $x < k/n = \lambda$ ist dieses positiv, für $x > k/n = \lambda$ ist dieses negativ. Also ist p_k für $0 < x < \lambda$ streng monoton steigend und für $\lambda < x < 1$ streng monoton fallend. Insbesondere hat p_k bei $x = \lambda$ einen Hochpunkt. In diesem gilt:

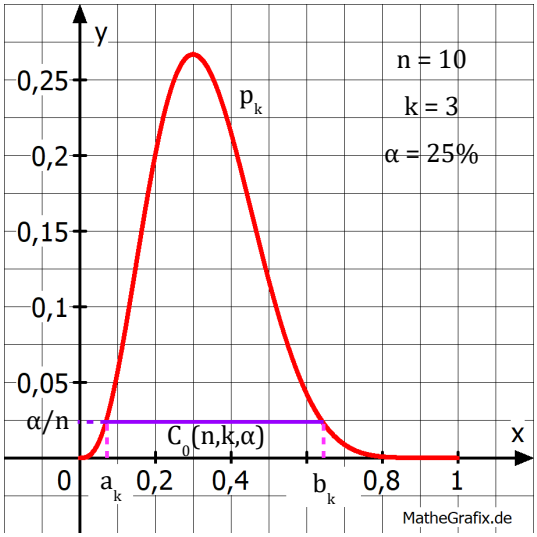
$$p_k(\lambda) = \binom{n}{k}\left(\frac{k}{n}\right)^k\left(\frac{n-k}{n}\right)^{n-k} = \binom{n}{k}\frac{k^k(n-k)^{n-k}}{n^n}$$

Die Rekursionsformeln lassen sich einfach nachrechnen und folgen aus den Eigenschaften der jeweiligen Binomialkoeffizienten:

$$\begin{aligned}
p_{k-1}(x) &= \binom{n}{k-1}x^{k-1}(1-x)^{n-k+1} \\
&= \frac{k}{n-k+1}\binom{n}{k}\cdot\frac{x^k}{x}\cdot(1-x)^{n-k}\cdot(1-x) \\
&= \frac{k}{n-k+1}\cdot\frac{1-x}{x}\cdot\binom{n}{k}x^k(1-x)^{n-k} \\
q_k(x) &= \binom{n+1}{k}x^k(1-x)^{n+1-k} \\
&= \frac{n+1}{n+1-k}\cdot\binom{n}{k}x^k(1-x)^{n-k}(1-x) \\
&= \frac{n+1}{n+1-k}\cdot(1-x)\cdot p_k(x) \\
&= \frac{n+1}{n+1-k}\cdot(1-x)\cdot\frac{n-k+1}{k}\cdot\frac{x}{1-x}\cdot p_{k-1}(x) \\
&= \frac{n+1}{k}x\cdot p_{k-1}(x)
\end{aligned}$$

□

Die folgende Abbildung zeigt das Schaubild von $p_3(x)$. Man erkennt den Hochpunkt bei $x = 0.3$ und die strenge Monotonie links und rechts davon:



Ist die Irrtumswahrscheinlichkeit $\alpha = 25\%$ vorgelegt, dann sieht man einen Bereich, in dem das Polynom die Wahrscheinlichkeit $\alpha/n = 2.5\%$ übersteigt.

Diesen Bereich nehmen wir als Konfidenzintervall $C_0(n,k,\alpha)$ bei $k=3$ Erfolgen unter $n=10$ Versuchen. Dass dieser Ansatz Erfolg hat, formuliert der folgende Satz.

Satz 14.2: Erster Versuch
Seien wieder $1 \leq n \in \mathbb{N}$ und $p_k(x) \in \mathbb{Z}[x]$ wie in obigem Satz definiert. Wir nehmen ferner an, $\alpha \in [0,1]$ ist *fein genug*, soll heißen für alle $k \in 1 \ldots n$ gilt

$$\frac{\alpha}{n} \quad \leq \quad p_k\left(\frac{k}{n+1}\right)$$

Schließlich bezeichnen wir für beliebiges $k \in 0 \ldots n$ das Intervall, über dem p_k einen Wert von α/n oder mehr annimmt, mit $C_0(n,k,\alpha)$ d.h.

$$C_0(n,k,\alpha) \quad := \quad p_k^{-1}\left(\left[\frac{\alpha}{n},1\right]\right) \cap [0,1] \quad = \quad \left\{ x \in [0,1] \;\middle|\; \frac{\alpha}{n} \leq p_k(x) \right\}$$

- Für $k=0$ bzw. $k=n$ lässt sich das Intervall $C_0(n,k,\alpha)$ explizit angeben

$$\begin{aligned} C_0(n,0,\alpha) &= [0,b_0] \\ C_0(n,n,\alpha) &= [a_n,1] \end{aligned}$$

  wobei $b_0 = 1-(\alpha/n)^{1/n}$ und $a_n = (\alpha/n)^{1/n}$ sind. Ist hingegen $0 < k < n$, dann gibt es ein eindeutig bestimmtes $a_k \in [0,k/n]$ so dass $p_k(a_k) = \alpha/n$ und ein eindeutig bestimmtes $b_k \in [k/n,1]$ so dass auch $p_k(b_k) = \alpha/n$ ist. Und für diese gilt dann

$$C_0(n,k,\alpha) \quad = \quad [a_k,b_k]$$

- Die linken und rechten Grenzen der eben definierten Intervalle bilden aufsteigende Ketten, d.h. es gelten

$$0 \;=\; a_0 \;\leq\; a_1 \;\leq\; a_2 \;\leq\; \ldots \;\leq\; a_n \;=\; (\alpha/n)^{1/n}$$

$$1-(\alpha/n)^{1/n} \;=\; b_0 \;\leq\; b_1 \;\leq\; b_2 \;\leq\; \ldots \;\leq\; b_n \;=\; 1$$

  Ist nun $0 < x < 1$ beliebig vorgelegt, dann bezeichnen wir die Menge der k mit $x \in C_0(n.k,\alpha)$ mit $K_0(x,n,\alpha)$, d.h. wir setzen $K_0(x,n,\alpha) := \{\, k \in 0 \ldots n \mid x \in C_0(n,k,\alpha) \,\}$. Diese ist ein Intervall in $\mathbb{Z}$, d.h. es gibt zwei Zahlen $0 \leq a \leq b \leq n$ mit

$$K_0(x,n,\alpha) \quad = \quad a \ldots b$$

- Wir haben also zu jedem $k \in 0 \ldots n$ ein Konfidenzintervall $C_0(n,k,\alpha) \subseteq [0,1]$ definiert. Tatsächlich ist die Irrtumswahrscheinlichkeit, dass x bei k Erfolgen nicht in $C_0(n,k,\alpha)$ liegt höchstens α, d.h.

$$i(x,n) \quad = \quad \sum_{x \notin C_0(n,k,\alpha)} b(x,n,k) \quad < \quad \alpha$$

Beweis:
Zunächst ist $C_0(n,k,\alpha) = [a_k, b_k]$ ein wohldefiniertes Intervall, wobei $p_k(a_k) = \alpha/n$ und $p_k(b_k) = \alpha/n$ mit $a_k < k/n < b_k$ sind. Dies alles ist klar, aufgrund der Eigenschaften von $p_k(x)$ aus dem vorangegangenem Satz.

Wegen $a_0 = 0$ ist $a_0 \leq a_1$ trivial. Und für $k \geq 1$ zeigen wir nun ebenso $a_{k-1} \leq a_k$. Da $p_{k-1}(x)$ ebenfalls bis $(k-1)/n$ streng monoton steigt, genügt es $p_{k-1}(a_{k-1}) \leq p_{k-1}(a_k)$ zu zeigen. Mit der Rekursionsformel gilt aber

$$p_{k-1}(a_k) \quad = \quad \frac{k}{n-k+1} \cdot \frac{1-a_k}{a_k} \cdot p_k(a_k)$$

Wegen $p_k(a_k) = \alpha/n = p_{k-1}(a_{k-1}) > 0$ lässt sich dies aus der Ungleichung kürzen, so dass wir nur noch $1 \leq k/(n-k+1) \cdot (1-a_k)/a_k$ zeigen müssen. Dies lässt sich weiter äquivalent umformen, zu $(n-k+1)a_k \leq k(1-a_k)$ und weiter zu $a_k \leq k/(n+1)$. Dies fogt aber aus der Annahme, dass α fein genug ist, wie folgende Argumentation zeigt:

Nach Voraussetzung ist $p_k(a_k) = \alpha/n \leq p_k(k/(n+1))$. Wegen $k/(n+1) < k/n$ liegt a_k also noch im Bereich der streng steigenden Monotonie von p_k und damit folgt tatsächlich $a_k \leq k/(n+1)$. Insgesamt bilden die a_k eine aufsteigende Kette.

Genauso findet man $b_{k-1} \leq b_k$, diese liegen im streng monoton fallendem Bereich von $p_{k-1}(x)$, so dass diese Ungleichung äquivalent ist, zu $p_{k-1}(b_{k-1}) \geq p_{k-1}(b_k)$. Mit der Rekursionsformel wird daraus

$$p_{k-1}(b_k) \quad = \quad \frac{k}{n-k+1} \cdot \frac{1-b_k}{b_k} \cdot p_k(b_k)$$

und kürzt man wieder $p_k(b_k) = \alpha/n = p_{k-1}(b_{k-1}) > 0$ so kann man dies wieder äquivalent umformen, zu $b_k \geq k/(n+1)$. Dies ist aber trivial, wegen $b_k \geq k/n > k/(n+1)$. Also bilden die b_k ebenfalls eine aufsteigende Kette.

Damit ist $K_0(n,k,\alpha)$ tatsächlich ein Intervall in den ganzen Zahlen, denn es ist ein Schnitt zweier solcher Intervalle:

$$K_0(n,k,\alpha) \quad = \quad \{\, k \in 0 \ldots n \mid a_k \leq x \,\} \cap \{\, k \in 0 \ldots n \mid x \leq b_k \,\}$$

Wir zeigen nun noch $a_k \leq b_{k-1}$ für beliebiges $k \in 1 \ldots n$. Dazu zeigen wir $p_{k-1}(a_k) \geq p_{k-1}(b_{k-1})$. Wir verwenden wieder die Rekursionsformel

$$p_{k-1}(a_k) \quad = \quad \frac{k}{n-k+1} \cdot \frac{1-a_k}{a_k} \cdot p_k(a_k)$$

Es lässt sich wieder $p_{k-1}(b_{k-1}) = \alpha/n p_k(a_k) > 0$ kürzen, so dass die Ungleichung äquivalent ist, zu $k/(n-k+1) \cdot (1-a_k)/a_k \geq 1$. Dies wurde aber oben schon bewiesen. Wäre also $a_k > b_{k-1}$, dann müsste (aufgrund der streng fallenden Monotonie von p_{k-1} hinter b_{k-1}) auch $p_{k-1}(a_k) < p_{k-1}(b_{k-1})$ sein - im Widerspruch zu dem was gerade gezeigt wurde. Anschaulich gesprochen bedeutet $a_k \leq b_{k-1}$ aber, dass sich die Intervalle $C_0(n,k-1,\alpha)$ und $C_0(n,k,\alpha)$ überlappen. D.h. die Intervalle überdecken $[0,1]$ und damit muss x in mindestens einem dieser Intervalle liegen - insbesondere ist $K_0(n,k-1,\alpha) \neq \emptyset$.

Damit können wir dann aber schließlich die Irrtumswahrscheinlichkeit abschätzen: Ist $x \notin C_0(n,k,\alpha)$ dann ist nach Definition $p_k(x) < \alpha/n$. Und da es höchstens n Intervalle $C_0(n,k,\alpha)$ gibt, in denen x *nicht* liegt, können wir weiterhin abschätzen

$$i(x,n) \;=\; \sum_{x \notin C_0(n,k,\alpha)} p_k(x) \;<\; n \cdot \frac{\alpha}{n} \;=\; \alpha$$

□

2. Versuch: Die exakte Lösung
Dieser Ansatz liefert zwar verlässliche Intervalle, doch sind diese viel größer als nötig, da wir die Zahl der Irrtümer $x \notin C_0(n,k,\alpha)$ nicht betrachtet haben. Zudem lassen sich die Konfidenzintervalle nicht leicht berechnen, da dazu polynomiale Gleichungen hohen Grades gelöst werden müssen. Die naheliegende Idee ist es nicht die Einzelwahrscheinlichkeiten zu kontrollieren, sondern gleich die kumulierte Binomialverteilung $B(x,n,k)$ selbst zurecht zu schneiden. Die konkrete Berechnung der Grenzen der Konfidenzintervalle kann mit Hilfe der inverse Betaverteilung erfolgen:

Satz 14.3: Clopper-Pearson
Sei $1 \leq n \in \mathbb{N}$ und eine zulässige Irrtumswahrscheinlichkeit $\alpha \in [0,1]$ vorgelegt. Zu $k \in 0 \ldots n$ definieren wir das *Konfidenzintervall* $C_0(n,k,\alpha) := [a_k, b_k]$ wobei a_k die Eigenschaft hat: $a_0 := 0$ und für $k > 0$

$$B\left(a_k, n, k-1\right) \;=\; 1 - \frac{\alpha}{2}$$

Die obere Grenze b_k wird definiert durch die Eigenschaft $b_n := 1$ bzw. für $k < n$ durch die Gleichung

$$B\left(b_k, n, k\right) \;=\; \frac{\alpha}{2}$$

Dann sind a_k und b_k wohldefiniert und für die Irrtumswahrscheinlichkeit der Konfidenzintervalle $C_0(n,k,\alpha)$ gilt für alle $x \in [0,1]$

$$i(x,n) \;=\; \sum_{x \notin C_0(n,k,\alpha)} b(x,n,k) \;<\; \alpha$$

Bemerkung 14.4:
In obigem Satz bezeichnet - wie immer - $B(x,n,k)$ die kumulierten Wahrscheinlichkeiten der Binomialverteilung zur Erfolgswahrscheinlichkeit x. Man beachte, dass damit $B(x,n,k-1) = p_x(\omega < k)$ die Wahrscheinlichkeit ist, echt weniger als k Erfolge zu beobachten. Analog ist $B(x,n,k) = p_x(\omega \leq k)$ die Wahrscheinlichkeit mindestens k Erfolge zu erzielen. Nach Satz (8.2.(vi)) lassen sich diese aber mit Hilfe der Beta-Verteilung ausdrücken

$$B(n,x,k) \;:=\; \sum_{i=0}^{k} b(n,x,i) \;=\; \beta(1-x, n-k, k+1)$$

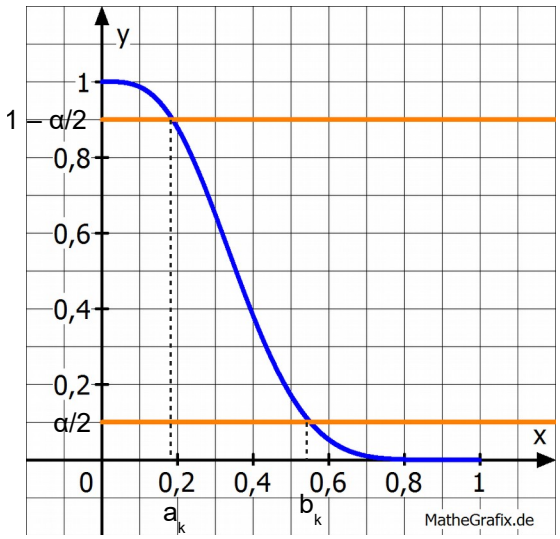
Diese Verteilung lässt sich nun aber (aufgrund der Stetigkeit und strengen Antitonie) invertieren. Die inverse Beta-Verteilung, also die Lösung x zu $y = \beta(x, a, b)$ bezeichnen wir mit $x = \beta^{-1}(y, a, b)$. Dann gilt für $0 < k < n$:

$$\begin{aligned} a_k &= 1 - \beta^{-1}\left(1 - \frac{\alpha}{2}, n - k + 1, k\right) \\ b_k &= 1 - \beta^{-1}\left(\frac{\alpha}{2}, n - k, k + 1\right) \end{aligned}$$

Bemerkung 14.5:
Die folgende Tabelle zeigt den Fortschritt von Clopper-Pearson gegenüber unserem ersten Versuch: Wir berechnen jeweils die Konfidenzintervalle bei $n = 10$ Versuchen bei einer zulässigen Irrtumswahrscheinlichkeit von $\alpha = 10\%$. Man sieht, dass die Intervalle viel kleiner sind

k	erster Versuch	Clopper-Pearson
0	[0.000, 0.369]	[0.000, 0.259]
1	[0.001, 0.499]	[0.005, 0.394]
2	[0.016, 0.604]	[0.037, 0.507]
3	[0.049, 0.694]	[0.087, 0.607]
4	[0.097, 0.774]	[0.150, 0.696]
5	[0.156, 0.844]	[0.222, 0.778]
6	[0.226, 0.903]	[0.304, 0.850]
7	[0.306, 0.951]	[0.393, 0.913]
8	[0.396, 0.984]	[0.493, 0.963]
9	[0.501, 0.999]	[0.606, 0.995]
10	[0.631, 1.000]	[0.741, 1.000]

Beweis:
Zunächst einmal ist klar, dass $B(x, n, k) = p_0(x) + ... + p_k(x)$ stetig in x ist. Ebenso ist $b(0, n, 0) = 1$ und $b(0, n, k) = 0$ für $k > 0$, so dass $B(0, n, k) = 1$ für alle $0 \leq k \leq n$ ist. Analog ist $b(1, n, n) = 1$ und $b(1, n, k) = 0$ für $k < n$, so dass $B(1, n, k) = 0$ für alle $0 \leq k < n$ ist. D.h. $B(x, n, k)$ fällt von $(0, 1)$ stetig auf $(1, 0)$. Die folgende Abbildung zeigt $B(x, 10, 3)$



Die Antitonie von $B(x,n,k)$ ist anschaulich klar (wenn die Erfolgswahrscheinlichkeit steigt, sinkt die Wahrscheinlichkeit nur k Erfolge zu haben), folgt aber auch sofort aus der expliziten Darstellung in (8.2.(vi)) als Integral. Damit lässt sich die Gleichung $B(x,n,k-1) = 1-\alpha/2$ für beliebiges α und alle $k > 0$ eindeutig lösen, so dass a_k wohldefiniert ist. Ebenso lässt sich die Gleichung $B(x,n,k) = \alpha/2$ für beliebiges α und alle $k < n$ eindeutig lösen, so dass b_k wohldefiniert ist. Und wegen $\alpha \leq 1$ ist $\alpha/2 \leq 1-\alpha/2$, so dass $a_k \leq b_k$ ist. Insgesamt ist $C_0(n,k,\alpha)$ für alle $k \in 0\ldots n$ wohldefiniert.

Wir haben nun also zu jedem $k \in 0\ldots n$ das Konfidenzintervall $C_k := C_0(n,k,\alpha) = [a_k, b_k]$ gegeben. Seien nun $i \leq j < n$, dann ist $B(x,n,j) = B(x,n,i) + b(x,n,i+1) + \cdots + b(x,n,j) \geq B(x,n,i)$. Insbesondere ist damit $B(b_i,n,j) \geq B(b_i,n,i) = \alpha/2$. Also hat die Gleichung $B(x,n,j) = \alpha/2$ noch keine Lösung im Intervall $[0,b_i]$ - die Lösung liegt erst rechts von b_i und das bedeutet $b_i \leq b_j$. Insgesamt haben wir also $b_0 \leq b_1 \leq \cdots \leq b_n = 1$ und mit demselben Argument sieht man auch $0 = a_0 \leq a_1 \leq \cdots \leq a_n$.

Sei nun $x \in [0,1]$ die tatsächliche, unbekannte Erfolgswahrscheinlichkeit des Versuches beliebig vorgelegt. Wir zeigen nun, dass die Irrtumswahrscheinlichkeit $i(x,n) \leq \alpha$ ist. Nach Konstruktion ist diese gegeben, als Summe über alle $b(x,n,k)$ wobei $x \notin C_k = [a_k, b_k]$ ist. Nun ist $x \notin [a_k, b_k]$ aber äquivalent zu $x < a_k$ oder $b_k < x$, also

$$i(x,n) \;=\; \sum_{x \notin C_k} b(x,n,k) \;=\; \sum_{x<a_k} b(x,n,k) + \sum_{x>b_k} b(x,n,k)$$

Nun bilden die a_k und b_k aber wie gesehen aufsteigende Ketten. Die Bedingung $x < a_k$ ist also erst ab einem gewissen Index t gegeben, bzw. die Bedingung $x > b_k$ kann nur bis zu einem gewissen Index s gelten. Es ist

$$i(x,n) \;=\; \sum_{k=t}^{n} b(x,n,k) + \sum_{k=0}^{s} b(x,n,k) \;=\; 1 - B(x,n,t-1) + B(x,n,s)$$

Dabei ist $t = \min\{\, k \in 0\ldots n \mid x < a_k \,\}$ und $s = \max\{\, k \in 0\ldots n \mid x > b_k \,\}$. Man beachte, dass für $x \geq a_n$ die Bedingung $x < a_k$ nie erfüllt wird, die erste Summe also komplett entfällt und für $x \leq b_0$ analog die Bedingung $x > b_k$ nie erfüllt wird, so dass dann die rechte Summe entfällt. In diesen beiden Fällen wird die Irrtumswahrscheinlichkeit also noch kleiner sein. Der kritischste Fall ist $b_0 < x < a_n$, d.h. wenn beide Summen vorkommen. Wenn wir diesen abschätzen können, sind die Fälle $x \geq a_n$ oder $x \leq b_0$ ebenfalls erledigt.

Nach Konstruktion ist gerade noch $b_s < x$, aufgrund der Antitonie bedeutet das aber $B(x,n,s) < B(b_s,n,s) = \alpha/2$. Ebenso ist gerade noch $x < a_t$ und damit $1-\alpha/2 = B(a_t,n,t-1) < B(x,n,t-1)$ woraus wiederum $1 - B(x,n,t-1) < \alpha/2$ wird. Insgesamt haben wir also

$$i(x,n) \;=\; 1 - B(x,n,t-1) + B(x,n,s) \;<\; \frac{\alpha}{2} + \frac{\alpha}{2} \;=\; \alpha$$

□

Clopper-Pearson ist die exakte Antwort auf das Problem - doch die Nachteile liegen auf der Hand: (1) die Berechnung des Konfidenzintervalls ist nur mit einem CAS oder einer Tabellenkalkulation durchzuführen.

Auch ist (2) die Berechnung leider nicht offensichtlich. Dazu kommt (3) für sehr kleine $x \approx 0$ oder sehr große $x \approx 1$ ist $i(x,n)$ deutlich kleiner als α, das Konfidenzintervall also zu groß. Wir suchen also nach einem Ansatz der einfacher zu handhaben ist und besser zugeschnittene Intervalle liefert, um den Preis, dass gelegentlich $i(x,n) > \alpha$ vorkommen kann.

3. Versuch: Der Ansatz von Wald

Eine Methode, die lange als Standard gegolten hat, betrachtet die Zufallsvariable ω auf $\mathcal{B}(n,x)$. Zunächst standardisieren wir ω, d.h. wir bilden

$$Z \;:=\; \frac{\omega - E(\omega)}{\sigma(\omega)} \;=\; \frac{\omega - nx}{\sqrt{nx(1-x)}}$$

Nach Moivre-Laplace geht die Binomialverteilung in die Normalverteilung über, und damit ist Z näherungsweise standard-normalverteilt. Ist die maximal zulässige Irrtumswahrscheinlichkeit α vorgelegt, dann können wir das Intervall $[-u,u]$ angeben, so dass Z die Häufigkeit $1-\alpha$ über $[-u,u]$ hat:

$$p_x(-u \leq Z \leq u) \;\approx\; 1-\alpha$$

Dieses u ist $u = \Phi^{-1}(1-\alpha/2)$ (siehe unten). Wir versuchen also die Menge $\{\, Z \in [-u,u] \,\}$ zu analysieren: $|Z| \leq u$ ist nach Konstruktion äquivalent, zu:

$$|\omega - nx| \;\leq\; u \cdot \sqrt{nx(1-x)}$$

In dieser Ungleichung ist der Parameter x ja leider unbekannt - aber er wird geschätzt, durch $\lambda := k/n$ wobei k die Auswertung (Ziehung) von ω war. Wir ersetzen also x durch λ und betrachten damit $|\omega - n\lambda| \leq u \cdot \sqrt{n\lambda(1-\lambda)}$. Durch Division mit n wird dieser Ausdruck zu

$$|\frac{\omega}{n} - \lambda| \;\leq\; u \cdot \sqrt{\frac{\lambda(1-\lambda)}{n}}$$

Nun ist ω/n aber ein Schätzer für den unbekannten Parameterwert x, wir ersetzen ω/n also durch x und finden damit Schranken für den Wert von x:

$$|x - \lambda| \;\leq\; u \cdot \sqrt{\frac{\lambda(1-\lambda)}{n}}$$

Diese Gleichung kann sofort gelöst werden und wir finden ein Lösungsintervall $C_0(n,k,\alpha) := [x_1, x_2]$, das als Konfidenzintervall (wir nennen es das *Konfidenzintervall nach Wald*) verwendet werden kann:

$$x_{1,2} \;:=\; \lambda \pm u \cdot \sqrt{\frac{\lambda(1-\lambda)}{n}}$$

wobei $\lambda = k/n$ und $u = \Phi^{-1}(1-\alpha/2)$. Dafür, dass bei der Herleitung des Intervalls so viele Ersetzungen stattgefunden haben, liefert es noch passable Ergebnisse. Der naheliegende Gedanke ist doch aber ω durch k zu ersetzen und die erste Ungleichung nach x aufzulösen:

4. Versuch: Aus der Tschebyscheff-Ungleichung
Wir erinnern uns an die Ungleichung von Tschebyscheff, die bemerkenswerter Weise ja für beliebige Zufallsvariablen X gilt. Sind $\mu = E(X)$ der Erwartungswert und $\sigma^2 = V(X)$ die Varianz von X, dann gilt für alle $u > 0$:

$$p\left(|X - \mu| \geq u\sigma\right) \quad \leq \quad \frac{1}{u^2}$$

Ist $\alpha = 1/u^2$ vorgelegt, dann kann man also sicher sein, dass die Werte, die X annimmt, mit einer Wahrscheinlichkeit von mindestens $1 - \alpha$ innerhalb des Intervalls $\mu \pm u\sigma$ liegen. Ist x wieder der unbekannte Parameter der Binomialverteilung, dann gilt für $X = \omega$ bekanntlich $\mu = nx$ und $\sigma = \sqrt{nx(1-x)}$. Wenn ω also den Wert k angenommen hat, dann haben wir also folgende Ungleichung zu lösen:

$$|k - nx| \quad \leq \quad u\sqrt{nx(1-x)}$$

Satz 14.6:
Seien $1 \leq n \in \mathbb{N}$ und $k \in 0 \ldots n$. Ferner seien $u > 0$ und $x \in [0, 1]$ beliebig vorgegeben. Dann besitzt die Ungleichung $|k - nx| \leq u\sqrt{nx(1-x)}$ das Lösungsintervall $[x_1, x_2]$ wobei

$$x_{1/2} \quad = \quad \frac{k + u^2/2}{n + u^2} \pm \frac{u}{2(n + u^2)}\sqrt{\frac{4k(n-k) + nu^2}{n}}$$

Beweis:
Da beide Seiten der Ungleichung positiv sind, ändert sich die Lösungsmenge nicht, wenn wir die Ungleichung quadrieren: $(k - nx)^2 \leq u^2nx(1-x)$. Es handelt sich also um eine quadratische Ungleichung

$$(k - nx)^2 - u^2nx(1-x) \quad \leq \quad 0$$

Diese Ungleichung lässt sich elementar lösen: $(k - nx)^2 = k^2 - 2knx + n^2x^2$ und $u^2nx(1-x) = nu^2x - nu^2x^2$, damit wird diese Ungleichung zu

$$(n^2 + nu^2)x^2 - (2kn + nu^2)x + k^2 \quad \leq \quad 0$$

$$n(n + u^2)x^2 - 2n(k + u^2/2)x + k^2 \quad \leq \quad 0$$

Bezeichnen wir $\widetilde{n} := n + u^2$ und $\widetilde{k} := k + u^2/2$ und $\widetilde{\lambda} := \widetilde{k}/\widetilde{n}$ und dividieren durch $n\widetilde{n}$, dann sind wir also bei folgender Ungleichung angelangt

$$x^2 - 2\widetilde{\lambda}x + \frac{k^2}{n\widetilde{n}} \quad \leq \quad 0$$

Wir haben es also mit einer nach oben geöffneten Parabel zu tun, von der wir den unter der x-Achse gelegenen Teil suchen. Die Lösungsmenge ist also das kompakte Intervall $[x_1, x_2]$, wobei x_1 und x_2 die beiden Lösungen der Gleichung $x^2 - 2\widetilde{\lambda}x + \frac{k^2}{n\widetilde{n}} = 0$ sind. Die Lösungsformel für quadratische Gleichungen liefert diese beiden Lösungen x_1 und x_2 also frei Haus, wir müssen sie nur noch auspacken:

$$x_{1,2} \;=\; \widetilde{\lambda} \pm \sqrt{\widetilde{\lambda}^2 - \frac{k^2}{n\widetilde{n}}}$$

$$\begin{aligned}
\widetilde{\lambda}^2 - \frac{k^2}{n\widetilde{n}} &= \left(\frac{k + u^2/2}{(n+u^2)}\right)^2 - \frac{k^2}{n(n+u^2)} \\
&= \frac{n(k^2 + ku^2 + u^4/4)}{n(n+u^2)^2} - \frac{nk^2 + k^2u^2)}{n(n+u^2)^2} \\
&= \frac{nk^2u^2 + nu^4/4 - k^2u^2}{n(n+u^2)^2} \\
&= \frac{u^2}{4n(n+u^2)^2}(4nk - 4k^2 + nu^2) \\
&= \frac{u^2}{4(n+u^2)^2} \cdot \frac{4k(n-k) + nu^2}{n}
\end{aligned}$$

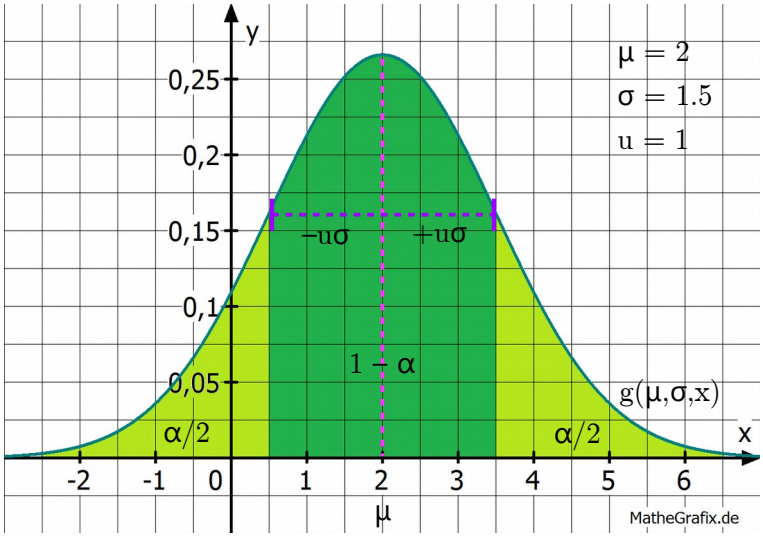
Zieht man aus $\widetilde{\lambda}^2 - k^2/(n\widetilde{n})$ die Wurzel, so kann man den vorderen Faktor also heraus ziehen, als $u/2(n+u^2)$ und damit erhält man genau die Lösungen x_1 und x_2, wie in der Behauptung gegeben.

□

Ist die Irrtumswahrscheinlichkeit $\alpha \in [0,1]$ vorgegeben, dann müssten wir nach Tschebyscheff $u = \sqrt{1/\alpha}$ wählen, um sicher zu gehen, dass $i(x,n) \leq \alpha$ ist. Das würde aber unsinnig große u ergeben, was daran liegt, dass die Ungleichung von Tschebyscheff für *jede beliebige* Verteilung von p gilt. Wir haben es hier aber mit der relativ zahmen Binomialverteilung zu tun. Eine exakte Abschätzung für u wäre möglich, aber Clopper-Pearson haben wir ja schon als zu kompliziert verworfen. Für große n ist die Binomialverteilung aber annähernd normalverteilt. Und für eine Normalverteilung lässt sich u deutlich leichter angeben, wie die folgende Bemerkung zeigt:

Bemerkung 14.7:
Wir bezeichnen die Verteilungsfunktion der Standardnormalverteilung mit $\Phi(u) := \Phi(0,1,u)$. Seien nun $\mu \in \mathbb{R}$ und $\sigma > 0$ beliebig fixiert und sei ferner ein $\alpha \in [0,1]$ vorgegeben. Dann ist das Intervall $A = [\mu - u\sigma, \mu + u\sigma] \subseteq \mathbb{R}$ für dass der Anteil $1-\alpha$ der (μ,σ)-normalverteilten Werte innerhalb A liegt (d.h. für das $\gamma(\mu,\sigma,A) = 1-\alpha$ gilt) gegeben, durch

$$u \;=\; \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)$$



Die folgende Tabelle vergleicht ein paar ausgewählte Werte für dieses u im Falle der Standardnormalverteilung ($\mu = 0$, $\sigma = 1$)

α	$\sqrt{1/\alpha}$	u
0.20	2.23607	1.28155
0.10	3.16228	1.64485
0.05	4.47214	1.95996
0.01	10.00000	2.57583

Beweis:
Es ist $\gamma(\mu, \sigma, A) = \Phi(\mu, \sigma, \mu + u\sigma) - \Phi(\mu, \sigma, \mu - u\sigma)$. Aufgrund der Punktsymmetrie zu $(\mu, 1/2)$ wird daraus aber $\gamma(\mu, \sigma, A) = 2(\Phi(\mu, \sigma, \mu + u\sigma) - 1/2) = 2\Phi(\mu, \sigma, \mu + u\sigma) - 1$. Nun nutzen wir noch die Identität $\Phi(\mu, \sigma, \mu + u\sigma) = \Phi(0, 1, u) = \Phi(u)$ aus und erhalten $1 - \alpha = \gamma(\mu, \sigma, A) = 2\Phi(u) - 1$ also $2\Phi(u) = 2 - \alpha$. Daraus wird dann sofort $u = \Phi^{-1}(1 - \alpha/2)$.

□

Definition 14.8: Konfidenzintervall nach Tschbyscheff
Seien $1 \leq n \in \mathbb{N}$ und $k \in 0 \ldots n$ und die zulässige Irrtumswahrscheinlichkeit $\alpha \in [0, 1]$ vorgelegt. Dann definieren wir das Konfidenzintervall $C_0(n, k, \alpha)$ zur Schätzung des Parameters x der Binomialverteilung $\mathcal{B}(n, x)$ bei $\omega = k$ Erfolgen, als $C_0(n, k, \alpha) := [x_1, x_2]$ wobei

$$\begin{aligned}
u &:= \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) \\
x_1 &:= \max\left\{0, \frac{k + u^2/2}{n + u^2} - \frac{u}{2(n + u^2)}\sqrt{\frac{4k(n-k) + nu^2}{n}}\right\} \\
x_2 &:= \min\left\{1, \frac{k + u^2/2}{n + u^2} + \frac{u}{2(n + u^2)}\sqrt{\frac{4k(n-k) + nu^2}{n}}\right\}
\end{aligned}$$

Lemma 14.9:
Seien α und $x \in [0,1]$ vorgelegt und bezeichne wieder $u := \Phi^{-1}(1-\alpha/2)$. Ferner sei $1 \leq n \in \mathbb{N}$ und $k \in 0 \ldots n$ fixiert. Wir bezeichnen schließlich noch $\mu := nx$ und $\sigma := \sqrt{nx(1-x)}$, dann sind äquivalent:

(a) $x \in C_0(n,k,\alpha) = [x_1, x_2]$ wie in (14.8)

(b) $k \in K_0(n,x,\alpha) := [\mu - u\sigma, \mu + u\sigma] \cap \mathbb{N}$

Beweis:
Wir bezeichnen die x_1 bzw. x_2 aus (14.6) mit y_1 und y_2. D.h. es sind $x_1 = \max\{0, y_1\}$ und $x_2 = \min\{1, y_2\}$. In (b) $\Longrightarrow$ (a) ist $k \in [\mu - u\sigma, \mu + u\sigma]$ gegeben. Es gilt also offenbar $|k - \mu| \leq u\sigma$ also $|k - nx| \leq u\sqrt{nx(1-x)}$. Nach (14.6) wissen wir damit aber $x \in [y_1, y_2]$ und $x \in [0,1]$ gilt bereits nach Voraussetzung, also auch $x \in [x_1, x_2]$.

In umgekehrter Richtung (a) $\Longrightarrow$ (b) ist $x \in [x_1, x_2] \subseteq [y_1, y_2]$ gegeben und nach (14.6) ist dies die Lösungsmenge der Ungleichung $|k - nx| \leq u\sqrt{nx(1-x)}$. Dies übersetzt sich aber wieder zu $|k - \mu| \leq u\sigma$ und da auch $k \in \mathbb{N}$ vorausgesetzt war, wird daraus $k \in K_0(n,x,\alpha)$.

□

Bemerkung 14.10:
Wenn wir die Lösungsformeln aus (14.6) haben, dann lässt sich die Formel von Wald sehr viel besser verstehen, als nur mit der direkten Herleitung in Versuch Nr. 3. Zunächst einmal stellen wir fest, dass statt n und k die verschobenen Ausdrücke $\widetilde{n} := n + u^2$ und $\widetilde{k} := k + u^2/2$ vorkommen. Und statt des Schätzers $\lambda = k/n$ tritt $\widetilde{\lambda} = \widetilde{k}/\widetilde{n}$ auf. Mit diesen Bezeichnungen wird die Lösungformel zu

$$x_{1/2} = \frac{\widetilde{k}}{\widetilde{n}} \pm \frac{u}{2\widetilde{n}}\sqrt{\frac{4k(n-k)+nu^2}{n}} = \widetilde{\lambda} \pm \frac{u}{2\widetilde{n}}\sqrt{\widetilde{\lambda}^2 - \frac{k^2}{n \cdot \widetilde{n}}}$$

Für große n und nicht zu kleine k bestehen dann die Näherungen $\widetilde{n} \approx n$, $\widetilde{k} \approx k$ und damit ist dann auch $\widetilde{\lambda} \approx \lambda$. Ferner können wir nu^2 gegenüber $4n(n-k)$ vernachlässigen. Damit geht diese Formel in die Wald-Formel über:

$$x_{1/2} \approx \lambda \pm \frac{u}{2n}\sqrt{\frac{4k(n-k)}{n}} = \lambda \pm u\sqrt{\frac{\lambda(1-\lambda)}{n}}$$

Bemerkung 14.11:
Es gibt noch einen zweiten, geometrischen Weg die Formel von Wald zu deuten. Wie in (14.6) beginnen wir wieder mit der Tschebyscheff-Ungleichung für die Binomialverteilung

$$|k - nx| \leq u\sqrt{nx(1-x)}$$

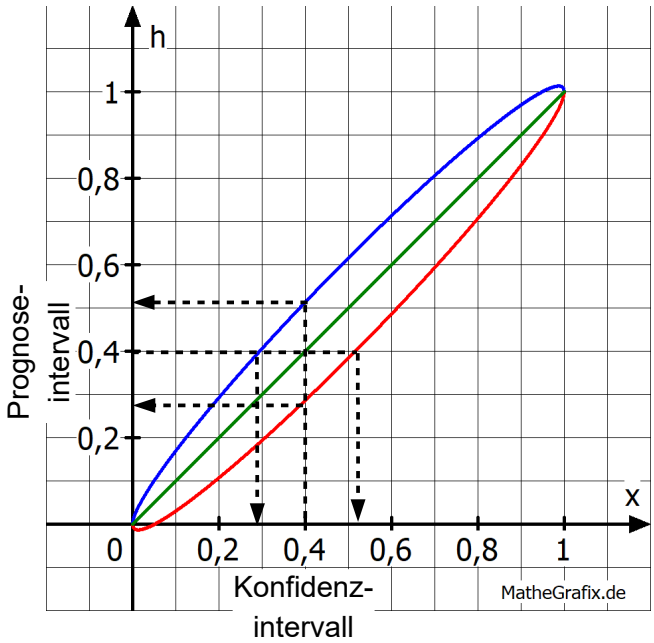
Kennen wir die Erfolgswahrscheinlichkeit x, dann können wir diese nach k auflösen und erhalten so ein Prognoseintervall $k_1 \ldots k_2$ für den Bereich, in dem die Zahl der Erfolge liegen wird - mit einer Irrtumswahrscheinlichkeit von $\alpha \approx 2(1 - \Phi(u))$, nämlich

$$k_{1/2} = nx \pm u\sqrt{nx(1-x)}$$

Durch Division durch n wird daraus ein Prognoseintervall für die relative Häufigkeit $h = k/n$ der Erfolge - mit derselben Irrtumswahrscheinlichkeit α

$$h_{1/2} \quad = \quad x \pm u\sqrt{\frac{x(1-x)}{n}}$$

Betrachtet man die Grenzen dieses Intervalls als Funktionen $h_{1/2}(x)$ in der Variablen x, dann erhält man ein Bild, das auch *Konfidenzellipse* genannt wird. Die folgende Abbildung zeigt diese für $n = 50$ und $u = 1.64$, also für eine Irrtumswahrscheinlichkeit von $\alpha = 10\%$



Liest man h_1 bzw. h_2 zu vorgegebenem x ab, so erhält man also das Prognoseintervall, in dem die relative Häufigkeit h mit $1-\alpha$ Wahrscheinlichkeit liegen wird. Dieses Intervall verwendet nur die Näherung der Binomialverteilung durch eine Normalverteilung um den Faktor u zu berechnen. Hier ist das Prognoseintervall zu $x = 40\%$ eingezeichnet.

Liest man aber umgekehrt zu vorgegebenem h die Werte $x_{1/2} = h_{1/2}^{-1}(h)$ ab, so erhält man das Konfidenzintervall nach Tschebyscheff zu $k = nh$. Dies hat dieselbe Wahrscheinlichkeit x zu überdecken von etwa $1-\alpha$, da es sich ja um eine äquivalente Umformulierung der Ungleichung handelt. Man sieht aber, dass die Ellipse *näherungsweise* symmetrisch zur grün eingezeichneten Winkelhalbierenden $h = x$ ist. Man kann also $x_{1/2}$ näherungsweise berechnen, indem man einfach x und h in der Formel für das Prognoseintervall vertauscht. Damit ergibt sich aber genau die Formel für das Konfidenzintervall nach Wald:

$$x_{1/2} \quad = \quad h \pm u\sqrt{\frac{h(1-h)}{n}}$$

Das Konfidenzintervall nach Tschebyscheff hat gegenüber Clopper-Pearson den klaren Vorteil einfach berechenbar zu sein, doch dafür haben wir einen hohen Preis bezahlt: indem wir u aus der Normalverteilung gewonnen haben, obwohl eine Binomialverteilung vorliegt, ist das Intervall nicht mehr exakt! Es gilt nicht mehr $i(x,n) \leq \alpha$ sondern nur noch $i(x,n) \approx \alpha$. Und $i(x,n)$ schwankt erstaunlich erratisch herum.
Es kann vorkommen, dass $i(x,n)$ sehr gut ist, aber $i(x,n+1)$ deutlich über α liegt. Näheres dazu findet sich in dem Übersichtsartikel [Brown, Cai, DasGupta; Interval Estimation for a binomial Proportion]. Wir wollen hier nur ein Beispiel angeben, für den häufigsten verwendeten Wert $\alpha = 0.05$:

$$\begin{aligned} i(0.5,25) &= 0.04329 \\ i(0.5,26) &= 0.07552 \end{aligned}$$

Wir sehen, dass sich die Irrtumswahrscheinlichkeit erhöht hat, obwohl durch die größere Zahl an Ziehungen die Annäherung an die Normalverteilung eigentlich besser sein sollte! Und das bei einem freundlichen Parameter von $x = 0.5$. Dieser Ansatz hat also leider noch Mängel.

Immerhin ist er schon ein deutlicher Fortschritt gegenüber dem Ansatz von Wald, dessen Abweichungen deutlich größer und genauso erratisch verteilt sind. Bevor wir die Mängel mildern, wollen wir aber zunächst einmal betrachten, wie man diesen Ausdruck für x_1 und x_2 zu verstehen hat und woher die Verbesserung gegenüber Wald kommt: Das Intervall hat nämlich nicht die Mitte $\lambda = k/n$, sondern ist verschoben worden, zu $\widetilde{\lambda} = \widetilde{k}/\widetilde{n}$ wobei $\widetilde{n} := n + u^2$ und $\widetilde{k} := k + u^2/2$. Dies liegt an der Asymmetrie der Binomialverteilung. Für $\alpha = 0.05$ ist $u \approx 2$ und damit $\widetilde{\lambda} \approx (k+2)/(n+4)$. Man kann dies interpretieren, als 4 weitere Ziehungen mit 2 weiteren Erfolgen. Der Wurzelausdruck lässt sich schlechter interpretieren, weswegen wir versucht sind eben diesen zu modifizieren.

5. Versuch: Ansatz von Agresti-Coull
Die folgende Modifikation stammt aus dem Jahr 1998 und hat sich aufgrund ihrer schönen Interpretierbarkeit und guten Ergebnisse neben Clopper-Pearson zum neuen Standard entwickelt

Definition 14.12: Konfidenzintervall nach Agresti-Coull
Seien $1 \leq n \in \mathbb{N}$ und $k \in 0 \ldots n$ und die zulässige Irrtumswahrscheinlichkeit $\alpha \in [0,1]$ vorgelegt. Dann definieren wir das Konfidenzintervall $C_0(n,k,\alpha)$ zur Schätzung des Parameters x der Binomialverteilung $\mathcal{B}(n,x)$ bei $\omega = k$ Erfolgen, als $C_0(n,k,\alpha) := [x_1, x_2]$ wobei

$$\begin{aligned} u &:= \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) \\ \widetilde{k} &:= k + u^2/2 \\ \widetilde{n} &:= n + u^2 \\ \widetilde{\lambda} &:= \frac{\widetilde{k}}{\widetilde{n}} \\ x_1 &:= \max\left\{ 0, \widetilde{\lambda} - u \cdot \sqrt{\frac{\widetilde{\lambda}(1-\widetilde{\lambda})}{\widetilde{n}}} \right\} \end{aligned}$$

$$x_2 \quad := \quad \min\left\{ 1, \widetilde{\lambda} + u \cdot \sqrt{\frac{\widetilde{\lambda}(1-\widetilde{\lambda})}{\widetilde{n}}} \right\}$$

Bemerkung 14.13:
Obwohl das Konfidenzintervall von Agresti-Coull eine schöne, durchsichtige Struktur hat, haben wir uns (insbesondere für große n) nur unwesentlich von dem Intervall nach Tschbyscheff wegbewegt. Wertet man diesen Ausdruck aus, so findet man letztlich nur einen marginalen Unterschied unter der Wurzel

$$\begin{aligned} x_1 \quad &:= \quad \max\left\{ 0, \frac{k+u^2/2}{n+u^2} - \frac{u}{2(n+u^2)}\sqrt{\frac{4k(n-k)+2nu^2+u^4}{n+u^2}} \right\} \\ x_2 \quad &:= \quad \min\left\{ 1, \frac{k+u^2/2}{n+u^2} + \frac{u}{2(n+u^2)}\sqrt{\frac{4k(n-k)+2nu^2+u^4}{n+u^2}} \right\} \end{aligned}$$

Konfidenzintervall zur Normalverteilung
Wir haben bisher immer Konfidenzintervalle einer binomial-verteilten Zufallsvariablen betrachtet. Wir wollen uns nun mit normal-verteilten Zufallsvariablen beschäftigen. Eine mögliche Problemstellung lautet also: Wir wollen Untersuchen ob Personen, die auf dem Land leben weniger intelligent sind, als Personen, die in der Stadt leben. Deswegen nehmen jeweils 20 zufällig ausgewählte Personen an einem Intelligenztest teil. Da die Auswahl aber zufällig war, genügt es nicht die mittleren IQs zu vergleichen, wir müssen Konfidenzintervalle bilden und prüfen, wie diese zueinander liegen.

Es geht also darum den Mittelwert μ einer normal-verteilten Größe zu ermitteln und einen Konfidenzbereich dafür anzugeben. Dabei müssen wir aber zwei verschiedene Situationen betrachten: (1) Im einfachsten Fall ist die Standardabweichung σ der Größe bekannt - beim Intelligenztest ist sie beispielsweise definiert. (2) In den meisten Fällen, müssen wir die Standardabweichung aber auch noch schätzen. Es wird sich zeigen, dass dieser Fall ist deutlich schwieriger ist:

Schätzung von μ bei bekanntem σ:
Sei also $X \in \mathcal{Z}(\mathbb{R})$ eine normal-verteilte Zufallsvariable, d.h. es ist $F_X = \Phi(\mu, \sigma)$, wobei μ unbekannt, aber σ bekannt ist. Wir legen ein Konfidenzniveau $\alpha \in [0,1]$ fest und suchen ein Schätzintervall $K_0(\alpha) \subseteq \mathbb{R}$ so dass $p(X \in K_0(\alpha)) \geq 1-\alpha$ ist. Dazu gehen wir wie folgt vor:

- Wir führen eine Stichprobe $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$ der Größe X aus. Als Schätzer für μ wählen wir das arithmetische Mittel $\lambda := A(x)$.

- Wie immer bezeichne Φ die Verteilungsfunktion der Standard Normalverteilung $\mathcal{N}(0,1)$, dann berechnen wir wieder $u := \Phi^{-1}(1-\alpha/2)$ und nehmen als Schätzintervall:

$$K_0(\alpha) \quad := \quad \left[\lambda - u\frac{\sigma}{\sqrt{n}},\ \lambda + u\frac{\sigma}{\sqrt{n}}\right]$$

Zur Begründung: Ist $X_k : S^n \to \mathbb{R} : (s_1, \ldots, s_n) \mapsto X(s_k)$ die k-te Ausführung der Ziehung von X und damit $A = (X_1 + \ldots + X_n)/n$ die Zufallsvariable des Mittels der n-fachen Durchführung von X, dann hat A den Erwartungswert $E(A) = \mu$ und die Varianz $V(A) = \sigma^2/n$. Damit ist A also $\Phi(\mu, \sigma/\sqrt{n})$-verteilt. Und folglich gilt für die Standardisierung von A:

$$p\big(A \in K_0(\alpha)\big) \;=\; p\left(-u \leq \frac{A-\mu}{\sigma/\sqrt{n}} \leq u\right) \;=\; 1-\alpha$$

Schätzung von μ und σ:
Zunächst benötigen wir speziell hierfür eine weitere Verteilung, die Studentsche T-Verteilung. Sie modelliert den Quotienten $\Phi/\sqrt{\chi_n/n}$ einer Standardnormal-verteilten Zufallsvariablen und einer davon unabhängigen Chi-Quadrat-verteilten Zufallsvariablen mit n Freiheitsgraden.

Definition 14.14:
Sei $1 \leq n \in \mathbb{N}$ festgelegt, wobei wir n als Zahl der *Freiheitsgrade* bezeichnen. Dann betrachten wir die folgende Dichtefunktion

$$t_n(x) \;:=\; \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\cdot\Gamma\left(\frac{n}{2}\right)}\left(1+\frac{x^2}{n}\right)^{-\frac{n+1}{2}}$$

Mit Hilfe dieser Dichtefunktion definieren wir das Wahrscheinlichkeitsmaß τ_n und dessen Verteilungsfunktion T_n. Wir nennen $\mathcal{T}(n) := (\mathbb{R}, \mathcal{B}(\mathbb{R}), \tau_n)$ die (Studentsche) T-**Verteilung**. Seien also $B \in \mathcal{B}(\mathbb{R})$ und $x \in \mathbb{R}$, dann:

$$\begin{aligned} \tau_n(B) \;&:=\; \int_B t_n(u)du \\ T_n(x) \;&:=\; \int_{-\infty}^{x} t_n(u)du \end{aligned}$$

Sei wieder $X \in \mathcal{Z}(\mathbb{R})$ eine normal-verteilte Zufallsvariable, d.h. es ist $F_X = \Phi(\mu, \sigma)$, wobei diesmal aber μ und σ unbekannt sind. Wir suchen wieder $K_0(\alpha) \subseteq \mathbb{R}$ so dass $p(X \in K_0(\alpha)) \geq 1-\alpha$ ist, für ein vorgelegtes $\alpha \in [0,1]$. Dazu gehen wir wie folgt vor:

- Wir führen eine Stichprobe $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$ der Größe X aus. Als Schätzer für μ wählen wir das arithmetische Mittel $\lambda := A(x)$.
- Aus dieser Stichprobe schätzen wir auch σ, der erwartungstreue Schätzer hierfür ist allerdings nicht $\sigma(x)$, sondern

$$\varrho \;:=\; \sqrt{\frac{n}{n-1}}\sigma(x) \;=\; \sqrt{\frac{1}{n-1}\sum_{k=1}^{n}(x_k-\lambda)^2}$$

- Sei T_{n-1} die Verteilungsfunktion der T-Verteilung $\mathcal{T}(n-1)$ zu $n-1$ Freiheitsgraden, dann berechnen wir $u := T_{n-1}^{-1}(1-\alpha/2)$ und nehmen als Konfidenzintervall:

$$K_0(\alpha) \;:=\; \left[\lambda - u\frac{\varrho}{\sqrt{n}},\; \lambda + u\frac{\varrho}{\sqrt{n}}\right]$$

Zur Begründung: Bezeichnen wir X_k und A wie oben, dann nehmen wir noch eine weitere Zufallsvariable auf $Q := (X_1 - A)^2 + \ldots + (X_n - A)^2$. Damit folgt Q/σ^2 einer χ^2-Verteilung mit $n-1$ Freiheitsgraden und es ist $E(Q/(n-1)) = \sigma^2$. Bezeichnen wir nun noch $S := \sqrt{Q/(n-1)}$ und $T := \sqrt{n} \cdot (A - \mu)/S$, dann kann man zeigen, dass T einer t-Verteilung mit $(n-1)$-Freiheitsgraden folgt. Und durch die Wahl von u gilt nun

$$p\big(A \in K_0(\alpha)\big) \;=\; p\,(-u \leq T \leq u) \;=\; 1 - \alpha$$

Statistische Modelle
An diesen beiden Beispielen sieht man bereits, dass es schwierig ist, ein einheitliches Konzept dafür anzugeben, wie man Konfidenzintervalle berechnen muss. Wir entwickeln zum Schluss des Abschnittes aber zumindest noch einen Formalismus, der geeignet ist, um das Problem der Konfidenzintervalle zu beschreiben.

Definition 14.15: Statistische Modelle:
Wir nennen $(S, \mathcal{A}, P)$ ein **statistisches Modell** über der **Parametermenge** Λ, falls $\mathcal{A}$ eine σ-Algebra auf der Menge S ist und $P = (p_\lambda)$ eine Familie (über $\lambda \in \Lambda$) von Wahrscheinlichkeitsmaßen $p_\lambda : \mathcal{A} \to [0,1]$ ist. Mit anderen Worten: zu jedem $\lambda \in \Lambda$ ist $(S, \mathcal{A}, p_\lambda)$ ein Wahrscheinlichkeitsraum.

Beispiel 14.16:
Im Grunde sind alle Wahrscheinlichkeitsverteilungen die wir bisher eingeführt haben statistische Modelle, da sie ja alle Parameter enthielten. Wir wollen nur zwei davon gesondert heraus stellen:

(i) Seien $1 \leq r$, $n \in \mathbb{N}$ und $\lambda = (\lambda_1, \ldots, \lambda_r) \in \Delta_r$ vorgelegt. Dabei bezeichnet Δ_r wie immer den $(r-1)$-dimensionalen *Standardsimplex*

$$\Delta_r \;=\; \{\, \lambda = (\lambda_1, \ldots, \lambda_r) \in \mathbb{R}^r \mid \lambda_i \geq 0, \lambda_1 + \cdots + \lambda_r = 1 \,\}$$

Dann ist die *Multinomialverteilung* $\mathcal{M}(n) := (S, \mathcal{A}, P)$ ein statistisches Modell über der Parametermenge $\Lambda = \Delta_r$, wobei $S = (1 \ldots r)^n$, $\mathcal{A} = \mathcal{P}(S)$ und $P = (p_\lambda)$ mit $\lambda \in \Delta_r$ ist. Dabei wurde die Wahrscheinlichkeitsverteilung $p_\lambda : S \to [0,1]$ bereits in (8.3) eingeführt. Offenbar modelliert $\mathcal{M}(n)$ die n-fache Ziehung mit Zurücklegen, aus einer Urne mir r Sorten von Kugeln.

(ii) Das statistische Modell der *Normalverteilung* $\mathcal{N} := (\mathbb{R}, \mathcal{B}(\mathbb{R}), \Gamma)$ hat die Wahrscheinlichkeitsmaße $\gamma(\mu, \sigma)$ mit Gauss-Funktion $g(\mu, \sigma)$ als Dichte - vergleiche (13.1). Die Parametermenge ist dabei $\Lambda := \mathbb{R} \times \mathbb{R}_{>0}$ d.h. es ist $\Gamma = (\gamma(\mu, \sigma))$ wobei $(\mu, \sigma) \in \Lambda$.

Bemerkung 14.17:

(i) Ist $(S, \mathcal{A}, P)$ ein statistisches Modell mit Parametermenge Λ und ist $X : S \to \mathbb{R}$ eine Zufallsvariable (d.h. für alle Borel-messbaren Mengen $B \in \mathcal{B}(\mathbb{R})$ gilt $X^{-1}(B) \in \mathcal{A}$), dann induziert X eine Familie $X_* P = (q_\lambda)$ von Wahrscheinlichkeitsmaßen auf $\mathbb{R}$, vermöge

$$q_\lambda \;:\; \mathcal{B}(\mathbb{R}) \to [0,1] \;:\; B \mapsto p_\lambda(X \in B)$$

D.h. $(\mathbb{R}, \mathcal{B}(\mathbb{R}), X_*P)$ ist wiederum ein statistisches Modell über derselben Parametermenge Λ. Und zu jedem $\lambda \in \Lambda$ bezeichnen wir die von q_λ induzierte Verteilungsfunktion mit

$$F_\lambda \ : \ \mathbb{R} \to [0,1] \ : \ x \mapsto p_\lambda(X \leq x)$$

(ii) Sei nun weiter T eine abzählbare Menge und $X : S \to T$ eine T-wertige Zufallsvariable auf S (d.h. für alle $t \in T$ gelte $X^{-1}(t) \in \mathcal{A}$). Ist nun $t \in T$ ein Zufallsereignis fixiert, dann definieren wir die Funktion

$$p_t \ : \ \Lambda \to [0,1] \ : \ \lambda \mapsto p_\lambda(X = t)$$

Ist nun zu jedem $t \in T$ ein $C_0(t) \subseteq \Lambda$ gegeben und bezeichnen wir $C_1(t) := \Lambda \setminus C_0(t)$, dann nennen wir die $C_0(t)$ eine Familie von *Konfidenzintervallen* zum *Konfidenzniveau* $1 - \alpha$, falls für alle $\lambda \in \Lambda$ gilt

$$i(\lambda, X) \ \leq \ \alpha$$

wobei $\alpha \in [0,1]$ vorgegeben ist. Und die *Irrtumswahrscheinlichkeit* $i(\lambda, X)$ der Familie $C_0(t)$ ist dabei definiert, durch

$$i(\lambda, X) \ := \ \sum_{t \in T} \begin{cases} p_t(\lambda) & \text{für } \lambda \in C_1(t) \\ 0 & \text{für } \lambda \in C_0(t) \end{cases}$$

(iii) Wie in (i) ist $q_\lambda : T \to [0,1] : t \mapsto p_t(\lambda)$ eine Wahrscheinlichkeitsverteilung auf T, d.h. (T, q_λ) ist ein Laplace-Raum. Der n-fache Produktraum $(T, q_\lambda)^{\otimes n}$ hat dann die Wahrscheinlichkeitsverteilung

$$q_\lambda^{\otimes n} \ : \ T^n \to [0,1] \ : \ (t_1, \ldots, t_n) \mapsto \prod_{i=1}^{n} p_\lambda(X = t_i)$$

Es bedeutet daher keine Einschränkung der Allgemeinheit, dass wir und auf ein Ergebnis $t \in T$ beschränkt haben. Ist $t = (t_1, \ldots, t_n) \in T^n$ eine Folge von Zufallsergebnissen, dann könnte man die Funktion

$$p_t \ : \ \Lambda \to [0,1] \ : \ \lambda \mapsto \prod_{i=1}^{n} p_\lambda(X = t_i)$$

betrachten. Damit ließe sich die Definition in (ii) wortgleich wiederholen. Dies erhält man aber auch, wenn man die n-fachen Produkträume unter der Zufallsvariablen $X^n : S^n \to T^n : (s_1, \ldots, s_n) \mapsto (X(s_1), \ldots, X(s_n))$ verwendet, liefert also nichts Neues.

Kapitel 15

Hypothesentests

Hypothesentests widmen sich folgender Fragestellung: Man beobachtet eine überraschende Abweichung von dem erwarteten Ergebnis und fragt sich, ob diese Abweichung rein zufällig war. Zum Beispiel: Man möchte meinen, ein Würfel liefert alle 6 Zahlen mit gleicher Wahrscheinlichkeit. Aber ist das noch haltbar, wenn bei 100 Würfen 25 Sechser geworfen wurden?

Zunächst muss also die Hypothese formuliert werden. Dabei soll die *Nullhypothese* H_0 ausdrücken, dass doch alles so ist, wie erwartet, die Abweichung also lediglich zufällig ist. In unserem Fall mit dem Würfel wäre also $H_0 : p(6) = 1/6$. Die *Alternativhypothese* H_1 ist normaler Weise die Verneinung von H_0. Wägt man aber zwei Hypothesen gegeneinander ab, so kann H_1 auch eine selbstständige Aussage sein. In unserem Fall wäre $H_1 : p(6) > 1/6$ angemessen.

Man legt also eine maximal zulässige *Irrtumswahrscheinlichkeit* $\alpha \in [0, 1]$ fest. Die Idee ist nun: man versucht H_1 (die Abweichung ist signifikant) zu rechtfertigen, indem man beweist, dass das Versuchsergebnis unter der Annahme von H_0 sehr unwahrscheinlich ($\leq \alpha$) ist. Wenn H_0 unwahrscheinlich ist (und es keine weiteren Alternativen gibt), dann ist H_1 also plausibel (mit einer Wahrscheinlichkeit von $\geq 1 - \alpha$). Wir nennen $1 - \alpha$ daher auch das *Signifikanzniveau* des Tests.

Weil es dabei aber um Wahrscheinlichkeiten geht, kann man sich auch täuschen. Grundsätzlich gibt es 4 mögliche Kombinationen:

	H_0 wurde bestätigt	H_0 wurde verworfen
H_0 ist wahr	alles richtg	α-Fehler
H_0 ist falsch	β-Fehler	alles richtig

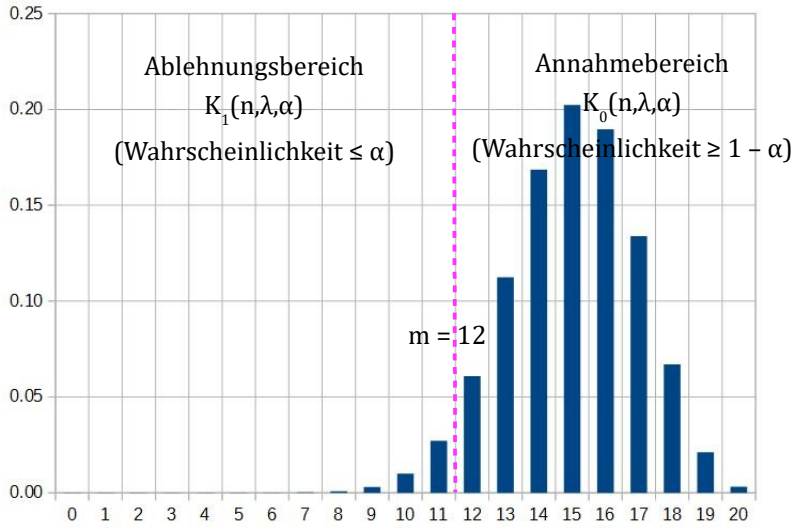
Ein α-Fehler (auch *Fehler 1. Art* genannt) liegt vor, wenn H_0 fälschlicher Weise verworfen wurde. Man wird (zumindest in diesem Test) nie erfahren, ob ein α-Fehler eingetreten ist, aber der Test wurde ja so konzipiert, dass die Wahrscheinlichkeit dafür höchstens α ist.

Ein β-Fehler (auch *Fehler 2. Art* genannt) liegt vor, wenn H_0 nicht verworfen wurde, obwohl sie unzutreffend ist. Die Wahrscheinlichkeit β dafür kann man (zumeist) nicht berechnen, wenn den zu testenden Parameter nicht kennt - und dann hätte man sich den Test ja sparen können. Dabei kann β sehr groß sein - zum Beispiel wenn der tatsächliche Wert des Parameters nahe (aber nicht genau auf) dem in H_0 vermuteten Wert liegt.

Beispiel 15.1:
Ein Großhändler für Feuerwerkskörper hat einen Dachschaden und zahlreiche Kisten sind feucht geworden. Der Lagerleiter schlägt vor, die Ware zum reduzierten Preis zu verkaufen und verspricht: *Mindestens* $\lambda = 75\%$ *der Ware ist noch voll funktionstüchtig.* Sie sollen testen, ob die Behauptung des Lagerleiters haltbar ist.

Sie entschließen sich also einen Test zu machen - es verbietet sich eine Vollerhebung durchzuführen, weil Sie dann nichts mehr zu verkaufen hätten. Sie nehmen sich also eine Kiste mit $n = 20$ Raketen und feuern diese ab. Es stellt sich heraus, dass nur $k = 13$ der Raketen starten - müssen Sie also die Behauptung des Lagerleiters verwerfen?

Wir wollen eine maximale Irrtumswahrscheinlichkeit von $\alpha = 0.05 = 5\%$ zulassen. Jede Rakete für sich kann einwandfrei funktionieren, oder eben nicht - es handelt sich also um eine Binomialverteilung. Die folgende Abbildung zeigt die Binomialverteilung $\mathcal{B}(20, 0.75)$:



Betrachtet man diese so stellt man fest, dass die Gesamtwahrscheinlichkeit der Zahlen 0 bis 11 funktionierender Raketen kanpp 5% beträgt. Die restlichen Anzahlen 12 bis 20 haben eine Wahrscheinlichkeit von etwas über 95%. Um mit einer Sicherheit von $1 - \alpha = 95\%$ sagen zu können, dass das Marketing die Lage zu optimistisch einschätzt, hätten also höchstens 11 Raketen nicht funktionieren dürfen. Da es aber 13 waren, kann die Hypothese der Marketing-Abteilung nicht verworfen werden. Wir formalisieren nun diese Vorgehensweise:

Test 15.2: Linksseitiger Test:
Wir gehen davon aus, dass die Zufallsvariable X binomialverteilt ist $\mathcal{B}(n, x)$, mit n Stufen und unbekannter Eintrittswahrscheinlichkeit x. Wir wollen die Hypothese $H_0 : x \geq \lambda$ für ein $\lambda \in [0, 1]$ testen. Dann gehe wie folgt vor:

1. Lege die maximal zulässige Irrtumswahrscheinlichkeit $\alpha \in [0, 1]$ fest. Berechne zu dieser den Maximalwert m, so dass

$$m \quad := \quad \max\left\{ k \in 0 \ldots n \mid \sum_{i=0}^{k} b(n, \lambda, i) \leq \alpha \right\}$$

Näherungsweise kann man die Binomialverteilung durch die Normalverteilung ersetzen, um die Berechnung von m zu vereinfachen. In diesem Fall sei $\mu := n\lambda$ und $\sigma^2 = n\lambda(1-\lambda) > 9$. Dann nehme

$$m \approx \sigma \cdot \Phi^{-1}(\alpha) + \mu - \frac{1}{2}$$

2. Damit definieren wir den *Ablehnungsbereich* $K_1(n, \lambda, \alpha) := 0 \ldots m$ und den *Annahmebereich* $K_0(n, \lambda, \alpha) := (m+1) \ldots n$.

3. Führe den Versuch X (ggf. alle n Wiederholungen eines einzelnen Versuchs X_1) aus, das Versuchsergebnis (also die Anzahl der Vorkommnisse in X) sei $k \in 0 \ldots n$.

4. Ist $k \in K_0(n, \lambda, \alpha)$ dann nimm die Hypothese H_0 an. Die Wahrscheinlichkeit β dafür, dass H_0 dennoch falsch ist, ist unbekannt und kann groß sein.

5. Ist $k \in K_1(n, \lambda, \alpha)$ dann verwerfe H_0 und nimm $H_1 : x < \lambda$ an. Die Wahrscheinlichkeit dafür, dass H_0 dennoch wahr ist, liegt höchstens bei α.

Test 15.3: Rechtsseitiger Test:
Wir gehen davon aus, dass die Zufallsvariable X binomialverteilt ist $\mathcal{B}(n, x)$, mit n Stufen und unbekannter Eintrittswahrscheinlichkeit x. Wir wollen die Hypothese $H_0 : x \leq \lambda$ für ein $\lambda \in [0, 1]$ testen. Dann gehe wie folgt vor:

1. Lege die maximal zulässige Irrtumswahrscheinlichkeit $\alpha \in [0, 1]$ fest. Berechne zu dieser den Minimalwert m, so dass

$$m := \min\left\{ k \in 0 \ldots n \mid 1 - \alpha \leq \sum_{i=0}^{k} b(n, \lambda, i) \right\}$$

Näherungsweise kann man die Binomialverteilung durch die Normalverteilung ersetzen, um die Berechnung von m zu vereinfachen. In diesem Fall sei $\mu := n\lambda$ und $\sigma^2 = n\lambda(1-\lambda) > 9$. Dann nehme

$$m \approx \sigma \cdot \Phi^{-1}(1-\alpha) + \mu - \frac{1}{2}$$

2. Damit definieren wir den *Annahmebereich* $K_0(n, \lambda, \alpha) := 0 \ldots m$ und den *Ablehnungsbereich* $K_1(n, \lambda, \alpha) := (m+1) \ldots n$.

3. Führe den Versuch X (ggf. alle n Wiederholungen eines einzelnen Versuchs X_1) aus, das Versuchsergebnis (also die Anzahl der Vorkommnisse in X) sei $k \in 0 \ldots n$.

4. Ist $k \in K_0(n, \lambda, \alpha)$ dann nimm die Hypothese H_0 an. Die Wahrscheinlichkeit β dafür, dass H_0 dennoch falsch ist, ist unbekannt und kann groß sein.

5. Ist $k \in K_1(n, \lambda, \alpha)$ dann verwerfe H_0 und nimm $H_1 : x > \lambda$ an. Die Wahrscheinlichkeit dafür, dass H_0 dennoch wahr ist, liegt höchstens bei α.

Beispiel 15.4:
Ein Würfel wird $n = 100$ Mal geworfen und zeigt $k = 25$ Mal eine Sechs an. Das kommt uns erstaunlich vor und deswegen fragen wir uns, ob das noch eine zufällige Abweichung sein kann. Die Nullhypothese lautet also, dass die unbekannte Wahrscheinlichkeit $x \in [0, 1]$ eine Sechs zu würfeln nur $x = 1/6$ beträgt. Kurz

$$H_0 \ : \ x \ = \ \frac{1}{6}$$

Die Verneinung von H_0 ist $H_1 : x \neq 1/6$, sollten wir also H_0 ablehnen, dann müssen wir H_1 annehmen. Als Irrtumswahrscheinlichkeit lassen wir wie üblich $\alpha = 0.05 = 5\%$ zu. Wie kann man entscheiden, ob H_0 zu verwerfen ist? In diesem Fall könnten wir natürlich Konfidenzintervalle verwenden: Wir berechnen mit der Methode von Agresti-Coull:

$$C_0(n, k, \alpha) \ \approx \ [0.175, 0.343]$$

Offenbar ist knapp $1/6 \notin C_0(n, k, \alpha)$, d.h. nur in höchstens $\alpha = 5\%$ der Fälle hätte ein Würfel mit $x = p(6) = 1/6$ eine Anzahl von $k = 25$ Sechsen bei $n = 100$ Würfen ergeben. Wir sind uns also $1 - \alpha = 95\%$ sicher, dass H_0 falsch ist. Daher nehmen wir H_1 an. Eine genauere Rechnung zeigt, dass in diesem Fall die Irrtumswahrscheinlichkeit 4.3% beträgt. Berechnet man das Konfidenzintervall aber nicht mit Agresti-Coull (14.12) sondern mit dem Ansatz aus Tschebyscheff (14.8), dann ergibt sich fast dasselbe Intervall (die Abweichung tritt bei einer Rundung auf 3 Nachkommastellen nicht auf) die genaue Irrtumswahrscheinlichkeit läge aber bei 5.9%.

Wie wir in (14.9) gesehen haben, ist der Ansatz mit dem Konfidenzintervall aus dem Ansatz von Tschebyscheff äquivalent dazu, dass k im Intervall $K_0(n, \lambda, \alpha)$ liegt. D.h. es gilt

$$x \in C_0(n, k, \alpha) \quad \Longleftrightarrow \quad k \in K_0(n, x, \alpha)$$

Dabei ist $K_0(n, x, \alpha) = [\mu - u\sigma, \mu + u\sigma] \cap \mathbb{N}$ mit $\mu = nx$, $\sigma = \sqrt{nx(1-x)}$ und $u = \Phi^{-1}(1 - \alpha/2)$. In unserem Fall ergibt sich also $K_0(100, 1/6, 5\%) = 10 \ldots 24$ und man sieht wieder, dass H_0 gerade noch abgelehnt werden muss. Auf diese Weise erspart man sich also die Berechnung des Konfidenzintervalls, weswegen wir das Schema des Punkttests auch so formulieren wollen:

Test 15.5: Punkttest: (Beidseitiger Test)
Wir gehen davon aus, dass die Zufallsvariable X binomialverteilt ist $\mathcal{B}(n, x)$, mit n Stufen und unbekannter Eintrittswahrscheinlichkeit x. Wir wollen die Hypothese $H_0 : x = \lambda$ für ein $\lambda \in [0, 1]$ testen. Dann gehe wie folgt vor:

1. Lege die maximal zulässige Irrtumswahrscheinlichkeit $\alpha \in [0, 0.5]$ fest. Zu dieser berechne a und $b \in 0 \ldots n$ so dass

$$a \ := \ \min\left\{ k \in 0 \ldots n \mid \frac{\alpha}{2} < \sum_{i=0}^{k} b(n, \lambda, i) \right\}$$

$$b \ := \ \min\left\{ k \in 0 \ldots n \mid 1 - \frac{\alpha}{2} \leq \sum_{i=0}^{k} b(n, \lambda, i) \right\}$$

Damit definieren wir den *Annahmebereich* $K_0(n, \lambda, \alpha) := a \dots b$ und den *Ablehnungsbereich* $K_1(n, \lambda, \alpha) := 0 \dots n \setminus K_0(n, \lambda, \alpha)$.

2. Näherungsweise kann man die Binomialverteilung durch den Ansatz $|k-\mu| \leq u \cdot \sigma$ ersetzen, um die Berechnung von a und b zu vereinfachen. In diesem Fall sei $\mu := n\lambda$ und $\sigma^2 := n\lambda(1-\lambda)$ und $u := \Phi^{-1}(1-\alpha/2)$. Dann nehme

$$K_0(n, \lambda, \alpha) \quad := \quad [\mu - u\sigma, \mu + u\sigma] \cap \mathbb{N}$$

3. Führe den Versuch X (ggf. alle n Wiederholungen eines einzelnen Versuchs X_1) aus, das Versuchsergebnis (also die Anzahl der Vorkommnisse in X) sei $k \in 0 \dots n$.

4. Ist $k \in K_0(n, \lambda, \alpha)$ dann nimm die Hypothese H_0 an. Die Wahrscheinlichkeit β dafür, dass H_0 dennoch falsch ist, ist unbekannt und kann groß sein.

5. Ist $k \notin K_0(n, \lambda, \alpha)$ dann verwerfe H_0 und nimm $H_1 : x \neq \lambda$ an. Die Wahrscheinlichkeit dafür, dass H_0 dennoch wahr ist, liegt höchstens bei α.

Würde man x kennen (was einen Test ad absurdum führt), könnte man auch die Wahrscheinlichkeit β berechnen, dass H_0 angenommen wird, obwohl $x \neq \lambda$ ist, also wenn $k \in K_0(n, \lambda, \alpha)$ ist. Die Wahrscheinlichkeit, dass k vorkommt, ist $p_x(\omega = k)$. Insgesamt also

$$\beta \quad = \quad \sum_{k=a}^{b} p_x(\omega = k)$$

Wir haben nun also die verschiedenen Fälle eines Hypothesentests einer binomial verteilten Zufallsvariablen durch dekliniert. Es wird Zeit einen formalen Rahmen so finden, in dem sich zeigen lässt, das die eben beschriebenen Tests tatsächlich eine Irrtumswahrscheinlichkeit von höchstens α für einen α-Fehler haben. Da es sich bei Tests um ein sehr allgemeines Problem handelt, ist auch die folgende Definition sehr allgemein gehalten. Sie sagt nur was ein Test ist und wann er gut ist, nicht wie man einen Test findet:

Definition 15.6: Hypothesentest:
Sei $(S, \mathcal{A}, P)$ ein statistisches Modell über der Parametermenge Λ, d.h. $P = (p_\lambda)$ mit $\lambda \in \Lambda$ ist eine Familie von Wahrscheinlichkeitsmaßen auf S. Dabei sei $x \in \Lambda$ der tatsächlich vorkommende, unbekannte Parameterwert. Sei ferner $X : S \to T$ eine Zufallsvariable in einen beliebigen Maßraum $(T, \mathcal{B})$.

(i) Eine nicht-leere Teilmenge $\emptyset \neq \Lambda_0 \subseteq \Lambda$ nennen wir eine **Hypothese** an $(S, \mathcal{A}, P)$. Die *Nullhypothese* ist die Aussage $H_0 = (x \in \Lambda_0)$, die *Alternative* ist die Aussage $H_1 = (x \notin \Lambda_0)$.

(ii) Eine nicht-leere, messbare Teilmenge $K_0 \subseteq T$ mit $K_0 \in \mathcal{B}$ nennen wir **Annahmebereich**, ihr Komplement $K_1 := T \setminus K_0$ heißt **Ablehnungsbereich** für H_0.

(iii) Seien nun eine Hypothese $\Lambda_0 \subseteq \Lambda$ und ein Annahmebereich $K_0 \subseteq T$ fixiert. Ein **Test** (Λ_0, K_0, t) besteht nun darin, dass wir ein Zufallsexperiment mit X ausführen - das Ergebnis sei $t \in T$. Die Hypothese H_0 wird **angenommen**, wenn $t \in K_0$ bzw. **abgelehnt**, wenn $t \in K_1$ ist.

(iv) Wir sprechen von einem α-*Fehler*, wenn $x \in \Lambda_0$ aber $t \in K_1$ ist. Die Wahrscheinlichkeit eines solchen Fehlers ist (für $x \in \Lambda_0$):

$$\alpha \;=\; p_x(X \in K_1)$$

(v) Wir sprechen von einem β-*Fehler*, wenn $x \notin \Lambda_0$ aber $t \in K_0$ ist. Die Wahrscheinlichkeit eines solchen Fehlers ist (für $x \notin \Lambda_0$):

$$\beta \;=\; p_x(X \in K_0)$$

(vi) Sei nun $\alpha \in [0,1]$ eine festgelegte, maximal zulässige *Irrtumswahrscheinlichkeit*, $\Lambda_0 \subseteq \Lambda$ sei eine Hypothese und $B \subseteq T$ ein Annahmebereich für H_0. Dann nennen wir B **α-zulässig**, falls gilt:

$$\forall\, \lambda \in \Lambda_0 \;:\; p_\lambda(X \notin B) \;\leq\; \alpha$$

Und wir nennen $C \subseteq T$ einen **α-optimalen** Annahmebereich für H_0 wenn (1) C ist ein α-zulässiger Annahmebereich für H_0 und (2) C hat eine minimale Wahrscheinlichkeit, einen β-Fehler zu begehen: D.h. für alle α-zulässigen $B \subseteq T$ gilt

$$\forall\, \lambda \in \Lambda \setminus \Lambda_0 \;:\; p_\lambda(X \in C) \;\leq\; p_\lambda(X \in B)$$

Bemerkung 15.7:
Hypothesentests werden insbesondere in psychologischen Gutachten und vor Gericht verwendet. Nach der Konstruktion dieser Tests kann man aber nur die Irrtumswahrscheinlichkeit für einen α-Fehler kontrollieren. Es muss H_0 also die *Unschuldsvermutung* sein: Die Wahrscheinlichkeit einen Unschuldigen zu verurteilen soll klein sein. Je kleiner man α macht, desto höher wird natürlich β - die Wahrscheinlichkeit, dass ein Schuldiger nicht verurteilt wird. Ein optimaler Test ist also solcher, der 'nur' Schuldige trifft (mit Irrtumswahrscheinlichkeit $\leq \alpha$), aber möglichst wenige Schuldige ungestraft davon kommen lässt (also minimale Irrtumswahrscheinlichkeit β hat).

Satz 15.8:
Sei $(S, \mathcal{A}, P) = \mathcal{B}(n)$ das statistische Modell der n-stufigen Binomialverteilung. Wir betrachten $X = \omega : S \to 0 \ldots n$ die Zufallsvariable der Zahl der Erfolge. Sei schließlich $\alpha \in [0,1]$ eine vorgelegte, maximale Irrtumswahrscheinlichkeit. Dann gilt:

(i) Ist $\Lambda_0 = [\lambda, 1]$ für ein $\lambda \in [0,1]$, dann ist der oben für linksseitige Tests definierte Annahmebereich $K_0(n, \lambda, \alpha)$ ein α-zulässiger Annahmebereich für H_0.

(ii) Ist $\Lambda_0 = [0, \lambda]$ für ein $\lambda \in [0,1]$, dann ist der oben für rechtsseitige Tests definierte Annahmebereich $K_0(n, \lambda, \alpha)$ ein α-zulässiger Annahmebereich für H_0.

(iii) Ist $\Lambda_0 = \{\, \lambda \,\}$ für ein $\lambda \in [0,1]$, dann ist der oben für Punkttests definierte Annahmebereich $K_0(n, \lambda, \alpha)$ ein α-zulässiger Annahmebereich für H_0.

Diese Annahmebereiche sind im allgemeinen aber nicht α-optimal! Dies liegt daran, dass die Verteilungsfunktion der Binomialverteilung in diskreten Schritten steigt. Man kann also oft noch einige Einträge aus K_0 streichen, ohne die Schwelle α zu übersteigen, reduziert dabei aber den β-Fehler.

Beweis:
Da $(S, \mathcal{A}, P)$ das statistische Modell der n-stufigen Binomialverteilung ist, gilt $p_\lambda(i) = b(n, \lambda, i)$. Die kumulierte Verteilung der ersten k Wahrscheinlichkeiten $b(i, \lambda, i)$ bezeichnen wir, mit

$$B(n, \lambda, k) \quad := \quad \sum_{i=0}^{k} b(n, \lambda, i)$$

Und nach (8.2.(vi)) lässt sich diese Verteilung auch berechnen, mit Hilfe der β-Verteilung, genauer gilt:

$$B(n, \lambda, k) \quad = \quad \frac{n!}{(n-k-1)! \cdot k!} \int_0^{1-\lambda} t^{n-k-1}(1-t)^k dt$$

Ist nun $0 \leq \kappa \leq \lambda \leq 1$ dann ist $1 - \lambda \leq 1 - \kappa$ und damit sieht man an der obigen Identität sofort $B(n, \lambda, k) \leq B(n.\kappa, k)$. Mit dieser Antitonie der Binomialverteilung sieht man dann weiter:

(i) Nach Konstruktion ist $K_0(n, \lambda, \alpha) = (m+1) \ldots n$ für das maximale m mit $B(n, \lambda, m) \leq \alpha$. Wir verabreden nun die Abkürzungen $K_0 := K_0(n, \lambda, \alpha)$ und $K_1 := T \setminus K_0 = 0 \ldots m$. Dann zeigen wir die α-Zulässigkeit von K_0: sei also $\kappa \in \Lambda_0$, d.h. $\kappa \geq \lambda$. Dann ist aufgrund der Antitonie $p_\kappa(\omega \in K_1) = B(n, \kappa, m) \leq B(n, \lambda, m) \leq \alpha$. Also ist insgesamt $p_\kappa(\omega \notin K_0) \leq \alpha$, was die α-Zulässigkeit von K_0 bedeutet.

(ii) Nach Konstruktion ist $K_0(n, \lambda, \alpha) = 0 \ldots m$ für das minimale m mit der Eigenschaft $1 - \alpha \leq B(n, \lambda, m)$. Wir verabreden nun die Abkürzung $K_0 := K_0(n, \lambda, \alpha)$ und zeigen die α-Zulässigkeit von K_0: sei also $\kappa \in \Lambda_0$, d.h. $\kappa \leq \lambda$. Dann ist aufgrund der Antitonie $p_\kappa(\omega \in K_0) = B(n, \kappa, m) \geq B(n, \lambda, m) \geq 1 - \alpha$. Also ist insgesamt $1 - p_\kappa(\omega \notin K_0) = p_\kappa(\omega \in K_0) \geq 1 - \alpha$ und damit $p_\kappa(\omega \notin K_1) \leq \alpha$, was die α-Zulässigkeit von K_0 bedeutet.

(iii) Nach Konstruktion ist $K_0(n, \lambda, \alpha) = a \ldots b$, wobei a und b minimal sind, mit den Eigenschaften $\alpha/2 < B(n, \lambda, a)$ respektive $1 - \alpha/2 \leq B(n, \lambda, b)$. Wir verabreden nun die Abkürzung $K_0 := K_0(n, \lambda, \alpha)$ und zeigen die α-Zulässigkeit von K_0: Wegen $\Lambda_0 = \{ \lambda \}$ ist nur zu zeigen $p_\lambda(\omega \notin K_0) \leq 0$. Wir unterscheiden nun zwei Fälle: Ist $a = 0$, dann ist $K_0 = 0 \ldots b$ und damit $p_\lambda(w \in K_0) = B(n, \lambda, b) \geq 1 - \alpha/2$. Daher gilt $p_\lambda(\omega \notin K_0) = 1 - p_\lambda(w \in K_0) \leq 1 - (1 - \alpha/2) = \alpha/2 \leq \alpha$, was zu zeigen war. Sei nun also $a \geq 1$, dann ist offenbar

$$p_\lambda(\omega \in K_0) \;=\; p_\lambda(\omega \leq b) - p_\lambda(\omega < a) \;=\; B(n, \lambda, b) - B(n, \lambda, a - 1)$$

Nach Konstruktion ist $1 - \alpha/2 \leq B(n, \lambda, b)$ und da a minimal gewählt wurde, ist auch $B(n, \alpha, a - 1) \leq \alpha/2$. Wir finden also $p_\lambda(\omega \in K_0) \geq (1 - \alpha/2) - \alpha/2 = 1 - \alpha$ und damit $p_\lambda(\omega \notin K_0) = 1 - p_\lambda(\omega \in K_0) \leq \alpha$, so dass auch dieser Fall abgeschlossen wurde.

□

Die wohl bekannteste Gruppe von Hypothesenstests sind die Chi-Quadrat-Tests. Und dafür gibt es einen guten Grund: Nach den zentralen Grenzwertsatz ist die unabhängige Wiederholung eines Zufallsexperiments immer näherungsweise normalverteilt. Und die Varianz von n normalverteilten Zufallsvariablen ist χ_{n-1}-verteilt. Die Varianz misst aber die Wahrscheinlichkeit der Abweichung und damit ist kann man Tests konzipieren, die zu vorgelegtem Sicherheitsniveau α prüfen, ob die Annahme H_0 verworfen werden muss. Wir stellen einen ersten solchen Test vor:

Test 15.9: χ^2-Test auf Verteilung
Wir betrachten eine Zufallsvariable $X : R \to S$ zwischen zwei beliebigen Wahrscheinlichkeitsräumen (R, Σ_R, p) und (S, Σ_S, q). Wir nehmen weiter an, dass wir $1 \leq n \in \mathbb{N}$ Stichproben $x = (x_1, \ldots, x_n) \in S^n$ von X gemacht haben. Ist p bekannt, dann ist auch die Wahrscheinlichkeitsverteilung

$$p_X \;:\; \Sigma_S \to [0, 1] \;:\; C \;\mapsto\; p(X \in C)$$

bekannt. Oftmals ist p aber unbekannt und man hat nur eine Vermutung darüber - z.B: man vermutet, dass p gleichverteilt oder normalverteilt ist. Die Frage ist also, ob die Vermutung von der Stichprobe gestützt wird, oder ob es zu unwahrscheinlich ist, dass das vermutete p zu dieser Stichprobe x geführt hätte.

In dieser Allgemeinheit kann man p mit einer endlichen Stichprobe natürlich nicht überprüfen. Wir müssen uns also ebenfalls mit einem endlichen S begnügen - d.h. wir wählen eine Partition $C_1, \ldots, C_m \in \Sigma_S$, sprich es gilt

$$S \;=\; \bigcup_{i=1}^{m} C_i \quad \text{und} \quad \forall\, i, j \in 1 \ldots m \;:\; C_i \cap C_j = \emptyset$$

Wir haben also eine Vermutung über die Wahrscheinlichkeit $p_i \in [0,1]$ mit der das Ereignis $X \in C_i$ eintritt und wollen diese testen. Offenbar muss $p_1 + \cdots + p_m = 1$ gelten, da die C_i ja eine Partition von S sind. Bezeichnen wir die tatsächliche Zahl der Stichproben, die in der Klasse C_i gelegen haben

$$n_i \;:=\; \#\{\, k \in 1 \ldots n \mid x_k \in C_i \,\}$$

dann lässt sich die zu testende Hypothese $p(X \in C_i) = p_i$ wegen des Gesetzes der großen Zahlen also äquivalent formulieren, in die Form

$$H_0 \;:\; \forall\, i \in 1 \ldots m \;:\; \lim_{n \to \infty} \frac{n_i}{n} \;=\; p_i$$

Es ist also $e_i = n \cdot p_i$ die (unter der Annahme von H_0) erwartete Anzahl von Stichproben in der Klasse C_i. Dazu betrachten wir die relative, quadratische Abweichung der tatsächlichen, von der erwarteten Häufigkeit, in Formeln

$$\Delta \;:=\; \sum_{i=1}^{m} \frac{(n_i - e_i)^2}{e_i}$$

Dabei ist die Zufallsvariable Δ annähernd χ^2-verteilt, mit $m-1$ Freiheitsgraden. Als Faustregel sagt man, die Annäherung ist gut genug, wenn gilt

$$\forall\, i \in 1 \ldots m \;:\; e_i \;\geq\; 5$$

Für $m = 2$ sollte sogar $e_i \geq 10$ sein. Der Test funktioniert nun also wie folgt: Man legt ein Sicherheitsniveau $\alpha \in [0,1]$ fest und führt die Stichprobe $x = (x_1, \ldots, x_n) \in S^n$ aus. Dabei ist die Anzahl n so zu wählen, dass gilt

$$\forall\, i \in 1 \ldots m \;:\; e_i \;=\; n \cdot p(X \in C_i) \;\geq\; 5$$

Wir nehmen die Hypothese (die Verteilung p auf der Partition $C_1, \ldots, C_m$) an, wenn $\Delta \leq \chi_{m-1}(1-\alpha)$ ist. Dabei ist χ_{m-1} die Verteilungsfunktion der χ^2-Verteilung, wie in (13.9) beschrieben. Ist hingegen $\Delta > \chi_{m-1}(1-\alpha)$, dann ist die beobachtete Abweichung Δ (unter der Annahme von p) unwahrscheinlicher, als α, so dass p verworfen werden muss.

Beispiel 15.10:
In einem Krankenhaus werden $n = 400$ Kinder geboren - 194 Mädchen und 206 Jungen. Wir fragen uns also: Liegt die Wahrscheinlichkeit, dass ein Mädchen geboren wird tatsächlich bei 50%? Wir modellieren also: $R = 1 \ldots 400$ die geborenen Kinder, $S = \mathbb{B}$ die möglichen Geschlechter, wobei wir $0 :=$ Mädchen und $1 :=$ Junge setzten. $X : 1 \ldots 400 \to \mathbb{B}$ ist also die Zufallsvariable, die jedem Kind sein Geschlecht zuordnet. Die Klassierung ist natürlich $C_0 := \{\,\text{Mädchen}\,\}$ und $C_1 ? \{\,\text{Junge}\,\}$. Dann lautet die Hypothese

$$H_0 \;:\; p(X = \text{Mädchen}) \;=\; 50\%$$

Als Sicherheitsniveau wählen wir $\alpha = 95\%$. Dann berechnen wir zunächst die Erwartungswerte $e_0 = n \cdot p(X = 0) = 200$ und $e_1 = n \cdot p(X = 0) = 200$. Die Stichprobe ergab $n_0 = 194$ und $n_1 = 206$ und damit berechnen wir

$$\Delta \;:=\; \frac{(n_0 - e_0)^2}{e_0} + \frac{(n_1 - e_1)^2}{e_1} \;=\; \frac{16}{200} + \frac{36}{200} \;=\; \frac{1}{4}$$

Auf der anderen Seite berechnen wir (z.B. mit LibreOffice, dort kann $\chi_f(x)$ mit der Funktion $\mathsf{CHIINV}(x, f)$ berechnet werden) $\chi_1(5\%) \approx 3.8415 \geq \Delta$. Und damit wird H_0 durch die Stichprobe *nicht* verworfen.

In der Tat ist es aber so, dass auf 100 Mädchen etwa 105 Jungen geboren werden, die Wahrscheinlichkeiten also nicht gleich zu sein scheinen. Eine so geringe Abweichung findet man aber erst bei sehr großen Stichproben: Behalten wir H_0 bei, nehmen aber eine Stichprobe mit $n_0 = 100/205 \cdot n$ und $n_1 = 105/205 \cdot n$ an, dann müsste zur Widerlegung von H_0 gelten

$$\chi_1(1 - \alpha) \;<\; \Delta \;=\; \frac{(n_0 - 0.5n)^2}{0.5n} + \frac{(n_1 - 0.5n)^2}{0.5n} \;=\; \frac{1}{1618} n$$

Wenn uns wieder ein Sicherheitsniveau von 95% genügt, dann müssten wir also mindestens eine Stichprobe von 6216 Kindern Umfang durchführen. Werden dabei höchstens 3032 Mädchen geboren, verwerfen wir die Annahme H_0.

Test 15.11: Fisher's G-Test auf Verteilung
Eine Alternative zum Chi-Quadrat-Test ist der G-Test von Fisher. Wir verwenden alle Bezeichnungen x, C_i, n_i, e_i usw. wie eben im Chi-Quadrat Test auf Verteilung eingeführt. Ist λ_i (für $i \in 1 \ldots m$) die Wahrscheinlichkeit, dass X in C_i liegt und $\lambda = (\lambda_1, \ldots, \lambda_m) \in \Delta_m$, dann ist

$$L(\lambda) \;:=\; \prod_{i=1}^{m} \lambda_i^{n_i}$$

die Wahrscheinlichkeit genau n_1 mal $X \in C_1$ zu ziehen und genau n_2 mal $X \in C_2$ zu ziehen, usw. so, wie es in x geschah. Der MLE-Schätzer von λ_i zu diesen Häufigkeiten ist $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_m) = \text{MLE}(X, x) \in \Delta_m$ wobei

$$\varepsilon_i \;:=\; \frac{n_i}{n}$$

[Eigentlich müsste man eine Multinomialverteilung betrachten, aber der Binomialkoeffizient ändert als Vorfaktor nichts an der Position des Schätzers ε]. Wir vermuten die Wahrscheinlichkeitsverteilung $p = (p_1, \ldots, p_m) \in \Delta_m$ für X auf der Partition (C_i). Die Nullhypothese H_0 lautet also wieder: $p_i = p(X \in C_i)$ für alle $i \in 1 \ldots m$. Um diese zu testen betrachten wir diesmal aber folgendes Maß für die Abweichung

$$G \;:=\; 2\ln\left(\frac{L(\varepsilon)}{L(p)}\right) \;=\; 2\sum_{i=1}^{m} n_i \ln\left(\frac{n_i}{e_i}\right)$$

Da ε die Likelihood-Funktion L maximiert gilt $L(p) \leq L(\varepsilon)$ und damit $G \geq 0$. Wieder folgt G für $n \to \infty$ einer Chi-Quadrat-Verteilung zu $m-1$ Freiheitsgraden, denn es gilt

$$G \;=\; \Delta + \mathcal{O}\left(\left(n_i - e_i\right)^3\right)$$

Der Test läuft also wie folgt ab: Wir wählen wieder ein Sicherheitsniveau α und entscheiden: Ist $G \leq \chi_{m-1}(1-\alpha)$, dann akzeptieren wir H_0, ist hingegen $G > \chi_{m-1}(1-\alpha)$, dann müssen wir H_0 verwerfen.

Beweis:

(i) Zunächst einmal ist $\varepsilon = \textsc{mle}(X, x)$, d.h. ε ist die Stelle, an der L maximal wird. Da der Logarithmus streng monoton steigend ist genügt es das Maximum von $\ell(\lambda)$ aufzusuchen.

$$\ell(\lambda) \;:=\; \ln\big(L(\lambda)\big) \;=\; \sum_{i=1}^{m} n_i \ln(\lambda_i)$$

Da $\lambda \in \Delta_m$ hängt ℓ nur von $m-1$ Variablen ab, z.B. ist $\lambda_1 = 1 - \lambda_2 - \cdots - \lambda_m$. Damit ℓ ein Maximum hat, muss der Gradient von $\nabla\ell = 0$ verschwinden. Für $j \in 2 \ldots m$ ist aber

$$0 \;=\; \partial_j \ell(\lambda) \;=\; -n_1 \frac{1}{1 - \lambda_2 - \cdots - \lambda_m} + n_j \frac{1}{\lambda_j}$$

Das bedeutet $n_j/\lambda_j = n_1/\lambda_1$ und diese Gleichung besteht nun für alle $j \in 1 \ldots m$. Daraus wird aber $n_j = n_1 \cdot \lambda_j/\lambda_1$ und summieren wir diese Gleichungen ergibt sich

$$n \;=\; \sum_{j=1}^{m} n_j \;=\; \sum_{j=1}^{m} n_1 \frac{\lambda_j}{\lambda_1} \;=\; \frac{n_1}{\lambda_1} \sum_{j=1}^{m} \lambda_j \;=\; \frac{n_1}{\lambda_1}$$

Daraus folgt $\lambda_1 = n_1/n$ und analog finden wir $\lambda_j = n_j/n$ für alle $j \in 1 \ldots m$. Also ist ε ein kritischer Punkt von L. Da L aber am Rand von Δ_m verschwindet, positiv ist und dies der einizige kritische Punkt von L ist, muss es sich um ein Maximum handeln.

(ii) Des Weiteren gilt die Gleichheit in der Definition von G, denn es ist $n_i/e_i = (n\varepsilon_i)/(np_i) = \varepsilon_i/p_i$ und damit

$$\ln\left(\frac{L(\varepsilon)}{L(p)}\right) = \ln\left(\prod_{i=1}^{m}\left(\frac{\varepsilon_i}{p_i}\right)^{n_i}\right) = \sum_{i=1}^{m} n_i \ln\left(\frac{\varepsilon_i}{p_i}\right) = \sum_{i=1}^{m} n_i \ln\left(\frac{n_i}{e_i}\right)$$

(iii) Zuletzt setzten wir noch $\delta_i := n_i - e_i$ und rechnen $G = \Delta + \mathcal{O}(\delta_i^3)$ nach. Zunächst einmal ist

$$\Delta = \sum_{i=1}^{m} \frac{(n_i - e_i)^2}{e_i} = \sum_{i=1}^{m} \frac{\delta_i^2}{e_i}$$

Es ist aber auch klar, dass $\delta_1 + \cdots + \delta_m = n - n = 0$ ist. Und bekanntlich gilt $\ln(1+t) = t - t^2/2 + \mathcal{O}(t^3)$. Damit erhalten wir dann insgesamt

$$\begin{aligned}
G &= 2\sum_{i=1}^{m} n_i \ln\left(\frac{n_i}{e_i}\right) \\
&= 2\sum_{i=1}^{m} \big(e_i + \delta_i\big) \ln\left(1 + \frac{\delta_i}{e_i}\right) \\
&= 2\sum_{i=1}^{m} \big(e_i + \delta_i\big) \left(\frac{\delta_i}{e_i} - \frac{1}{2}\frac{\delta_i^2}{e_i^2} + \mathcal{O}\big(\delta_i^3\big)\right) \\
&= 2\sum_{i=1}^{m} \delta_i + 2\sum_{i=1}^{m} \frac{\delta_i^2}{e_i} - \sum_{i=1}^{m} \frac{\delta_i^2}{e_i} - \sum_{i=1}^{m} \frac{\delta_i^3}{e_i^2} + \mathcal{O}\big(\delta_i^3\big) \\
&= 0 + \sum_{i=1}^{m} \frac{\delta_i^2}{e_i} + \mathcal{O}\big(\delta_i^3\big) = \Delta + \mathcal{O}\big(\delta_i^3\big)
\end{aligned}$$

□

Beim Chi-Quadrat-Test haben wir also eine Vermutung über die Wahrscheinlichkeiten $p(X \in C_i)$ und will prüfen, ob diese Vermutung stimmt. Einen etwas anderen Ansatz verfolgt der Kolmogorov-Smirnov-Test den wir nur vorstellen. Hierbei haben wir eine Vermutung über die Verteilungsfunktion von X, die wir testen wollen. Dieser Test umgeht damit das Problem der Wahl der Klassen C_i, funktioniert aber nur bei reellen Zufallsvariablen. Für alle Beweise der Aussagen siehe die Open Courseware Notes des MIT zur Statistik, Section 13.

Test 15.12: Kolmogorov-Smirnov-Test auf Verteilung
Sei (S, Σ, p) ein Wahrscheinlichkeitsraum und $X : S \to \mathbb{R}$ eine reelle Zufallsvariable darauf. Die tatsächliche, unbekannte Verteilungsfunktion von X sei F_X also $F_X : \mathbb{R} \to [0,1] : t \mapsto p(X \leq t)$. Wir vermuten, dass X die stetige Verteilungsfunktion $F_0 : \mathbb{R} \to [0,1]$ hat, d.h. unsere Hypothese lautet

$$H_0 \;:\; F_X \;=\; F_0$$

Um H_0 zu testen führen wir eine Stichprobe $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$ aus und wir bezeichnen die *empirische Verteilungsfunktion* von x mit

$$F_x(t) \;:=\; \frac{1}{n} \# \{\, k \in 1 \ldots n \mid x_k \leq t \,\}$$

Nach Glienko-Cantelli gilt $(F_x) \to F_X$ p-fast sicher gleichmäßig, wenn die Anzahl der Stichproben $n \to \infty$ wächst. Unter der Hypothese H_0 gilt also $(F_x) \to F_0$ und daher definieren wir die Abweichung

$$\Delta_n \;:=\; \sqrt{n} \cdot \|F_x - F_0\|_\infty$$

Sortieren wir x aufsteigend, d.h. $x_1 \leq x_2 \leq \cdots \leq x_n$ dann folgt aus der Tatsache, dass F_x stückweise konstant und F_0 stetig ist, dass

$$\begin{aligned}
\Delta_n^{\text{hi}} &:= \max\{\, |F_x(x_k) - F_0(x)| \mid k \in 1 \ldots n \,\} \\
\Delta_n^{\text{lo}} &:= \max\{\, |F_x(x_{k-1}) - F_0(x)| \mid k \in 2 \ldots n \,\} \\
\|F_x - F_0\|_\infty &= \max\{\, \Delta_n^{\text{hi}}, \Delta_n^{\text{lo}}, F_0(x_1) \,\}
\end{aligned}$$

Ist F_X stetig, dann ist die Verteilungsfunktion von $\|F_x - F_X\|_\infty$ unabhängig von F_X. Und damit lässt sich dann zeigen, dass

$$\lim_{n \to \infty} p\big(\Delta_n \leq t\big) \;=\; H(t)$$

wobei $H(t)$ die Kolmogorov-Verteilung ist (diese lässt sich zwar nicht mit LibreOffice berechnen, aber es gibt die Möglichkeit dazu auf auf der Seite *www.math.ucla.edu/~tom/distributions/Kolmogorov.html*)

$$H(t) \;=\; 1 + 2 \sum_{k=1}^{\infty} (-1)^k e^{-2k^2 t}$$

Der Test funktioniert nun also wie folgt: Wähle ein Sicherheitsniveau α und c so, dass $H(c) = 1 - \alpha$, also $c = H^{-1}(1 - \alpha)$. Ist nun $\Delta_n \leq c$, dann akzeptieren wir H_0, ist hingegen $\Delta_n > c$, dann muss H_0 verworfen werden. Für $n \leq 35$ liegen die Werte von c auch tabelliert vor, für $n > 35$ kann man folgende Näherung verwenden:

$$c \;\approx\; \sqrt{\frac{1}{2} \ln\left(\frac{2}{\alpha}\right)}$$

Test 15.13: 8 **Kolmogorov-Smirnov-Test auf gleiche Verteilung**
Dieser Test ähnelt dem voran gegangen Test auf Verteilung. Diesmal lautet die Fragestellung aber: *Haben die Zufallsvariablen X und Y dieselbe Verteilung?* ohne dabei festlegen zu wollen, welche Verteilung dies sein sollte. Wir befinden uns also in folgender Situation:

Sei (S, Σ, p) ein Wahrscheinlichkeitsraum und X, $Y : S \to \mathbb{R}$ zwei reelle Zufallsvariable darauf. Die tatsächlichen, unbekannten Verteilungsfunktionen von seien F_X von X bzw. F_Y von Y. Wir führen nun zwei Stichproben $x = (x_1, \ldots, x_m) \in \mathbb{R}^m$ und $y = (y_1, \ldots, y_n) \in \mathbb{R}^n$ aus. Mit F_x bzw. F_y bezeichnen wir wieder die empirischen Verteilungsfunktionen zu diesen Stichproben. Dann berechnen wir

$$\Delta_{m,n} \quad := \quad \sqrt{\frac{mn}{m+n}} \cdot \|F_x - F_y\|_\infty$$

Als Faustregel gilt: Der Test liefert gute Ergebnisse, wenn $mn \geq 4(m+n)$ ist, also z.B. für $m \geq 8$ und $n \geq 8$. Ansonsten ist alles wie zuvor: Wir legen ein Sicherheitsniveau α fest und berechnen $c = H^{-1}(1-\alpha)$. Ist $\Delta_{m,n} \leq c$, dann können wir $F_X = F_Y$ annehmen, ist $\Delta_{m,n} > c$, dann müssen wir die Hypothese $F_X = F_Y$ verwerfen.

Kapitel 16

Unabhängigkeitstests

Wir haben für zwei Ereignisse $A, B \subseteq S$ definiert, dass A und B *unabhängig* sind, wenn $p(A \cap B) = p(A) \cdot p(B)$. Wir betrachten nun zwei endliche Familien von Ereignissen $A_1, \ldots, A_m \subseteq S$ und $B_1, \ldots, B_n \subseteq S$. Dann ist die Frage, ob jedes einzelne A_i von jedem einzelnen B_j unabhängig ist - wir haben das bereits als Unabhängigkeit der Familien $\mathcal{A} := \{ A_1, \ldots, A_m \}$ und $\mathcal{B} := \{ B_1, \ldots, B_n \}$ kennen gelernt.

Ein einfaches Beispiel ist die Gegenüberstellung von Ausbildung und Arbeitsverhältnis, wie sie zu Beginn von Kapitel 5 gegeben wurde. Hier sind diese beiden Eigenschaften offensichtlich nicht unabhängig. Wir wollen diesen häufig vorkommenden Spezialfall der Definition (12.1) von zwei Zufallsvariablen mit endlichem Wertebereich, nun eingehender untersuchen:

Bemerkung 16.1:
Sei (S, Σ, p) ein Wahrscheinlichkeitsraum und $1 \leq m$, $n \in \mathbb{N}$ gegeben. Sind nun $X : S \to 1 \ldots m$ und $Y : S \to 1 \ldots n$ zwei Zufallsvariablen mit endlichem Wertebereich. Dann sind X und Y nach (12.1) **unabhängig**, falls für alle $i \in 1 \ldots m$ und alle $j \in 1 \ldots n$ gilt

$$p(X = i \text{ und } Y = j) \quad = \quad p(X = i) \cdot p(Y = j)$$

Bezeichnen wir $A_i := X^{-1}(i)$ und $B_j := Y^{-1}(j) \in \Sigma$, dann ist diese Gleichung anderes als eine suggestive Schreibweise für $p(A_i \cap B_j) = p(A_i) \cdot p(B_j)$. Und bezeichnen wir weiterhin $u_i := p(X = i) = p(A_i)$ und $v_j := p(Y = j) = p(B_j)$ und $p_{i,j} := p(A_i \cap B_j)$, dann lässt sich die Unabhängigkeit von X und Y abermals umschreiben, zu

$$\forall\, i \in 1 \ldots m \; \forall\, j \in 1 \ldots n \quad : \quad p_{i,j} \;=\; u_i \cdot v_j$$

In diesem Abschnitt widmen wir uns der Frage: Was tun, wir mit realen Daten, wenn $p(A_i \cap B_j)$ und $p(A_i) \cdot p(B_j)$ nur näherungsweise überein stimmen? Zunächst gehen wir auf die beste Schätzung der Wahrscheinlichkeiten $p(A_i)$ und $p(B_j)$ ein und danach auf Hypothesentests mit der Frage nach der Unabhängigkeit.

Lemma 16.2:
Sind $1 \leq m,\ n \in \mathbb{N}$ und ist $H = (h_{i,j}) \in \mathrm{mat}_{m\times n}(R)$ eine $m \times n$-Matrix (wobei R ein beliebiger Integritätsbereich sein kann, etwa $R = \mathbb{R}$) so dass die Summe aller Einträge der Matrix gleich 1 ist

$$\sum_{i=1}^{m}\sum_{j=1}^{n} h_{i,j} \quad = \quad 1$$

Dann bezeichnen wir $h_{+,j} := h_{1,j} + \cdots + h_{m,j}$ und $h_{i,+} := h_{i,1} + \cdots + h_{i,n} \in R$. Und damit sind dann die folgenden vier Aussagen äquivalent

(a) $\mathrm{rank}(H) = 1$

(b) $\forall\, i \in 1 \ldots m$ und $\forall\, j \in 1 \ldots n$ gilt $h_{i,j} = h_{i,+} \cdot h_{+,j}$

(c) es gibt Tupel $u = (u_1, u_2, \ldots, u_m) \in R^m$ und $v = (v_1, v_2, \ldots, v_n) \in R^n$ so dass $\forall\, i \in 1 \ldots m$ und $\forall\, k \in 1 \ldots n$ gilt $h_{i,k} = u_i \cdot v_k$

(d) $\forall\, i, k \in 1 \ldots m$ und $\forall\, j, \ell \in 1 \ldots n$ gilt $h_{i,\ell} \cdot h_{k,j} = h_{i,j} \cdot h_{k,\ell}$

Der Nutzen des Lemmas liegt auf der Hand: Um zu überprüfen, ob eine Kontingenztabelle $(h_{i,j})$ die Kombination zweier unabhängiger Eigenschaften X und Y ist, genügt es zu testen, ob $h_{i,\ell} \cdot h_{k,j} = h_{i,j} \cdot h_{k,\ell}$ für alle Indices $i, k \in 1 \ldots m$ und $j, \ell \in 1 \ldots n$ ist. Wenn ja, dann sind die $h_{i,j}$ einfach die unabhängige Kombination (also das Produkt) der Randhäufigkeiten $h_{+,j} = h_{1,j} + \cdots + h_{m,j}$ und $h_{i,+} = h_{i,1} + \cdots + h_{i,n}$.

Das Problem ist nur: Bei realen Daten werden die $h_{i,j}$ immer leichte Abweichungen von den zugrunde liegende Wahrscheinlichkeiten $u_i = p(A_i)$ bzw. $v_j = p(B_j)$ haben. In diesem Fall müssen wir also einen möglichst guten Schätzer für u_i und v_j finden. Es ist anzunehmen, dass der beste Schätzer $u_i = h_{i,+}$ und $v_j = h_{+,j}$ sein sollte. In der Tat gilt der folgende Satz:

Satz 16.3:
Zu jedem Paar (i, j) mit $i \in 1 \ldots n$ und $j \in 1 \ldots n$ sei $n_{i,j} \in \mathbb{N}$ die Anzahl der Vorkommnisse des zugehörigen Merkmalspaares unter den insgesamt

$$N \quad := \quad \sum_{i=1}^{m}\sum_{j=1}^{n} n_{i,j}$$

Merkmalsträgern. D.h. $h_{i,j} = n_{i,j}/N$ ist die relative Häufigkeit des Merkmalspaares zu (i, j). Dann definieren wir die Likelihood Funktion L für $u \in \Delta_m$ und $v \in \Delta_n$ durch

$$L(u, v) \quad := \quad \prod_{i=1}^{m}\prod_{j=1}^{n} u_i^{n_{i,j}} v_i^{n_{i,j}} \quad = \quad \left(\prod_{i=1}^{m} u_i^{n_{i,+}}\right) \cdot \left(\prod_{j=1}^{n} v_i^{n_{+,j}}\right)$$

Dabei ist wieder $n_{+,j} = n_{1,j} + \cdots + n_{m,j}$ und $n_{i,+} = n_{i,1} + \cdots + n_{i,n}$. Dann ist der MLE-Schätzer (u, v) von L (also das Maximum von L) gegeben, durch

$$u_i \;=\; \frac{n_{i,+}}{N} \quad \text{und} \quad v_j \;=\; \frac{n_{+,j}}{N}$$

Beweis:
Wegen der strengen Monotonie des Logarithmus genügt es die Funktion $\ell(u,v) = \ln L(u,v)$ zu maximieren. Diese ist gegeben, durch

$$\ell(u,v) \;=\; \sum_{i=1}^{m} n_{i,+} \ln(u_i) + \sum_{j=1}^{n} n_{+,j} \ln(v_i)$$

$$\frac{\partial}{\partial u_i}\ell(u,v) \;=\; n_{i,+}\frac{1}{u_i} \quad \text{und} \quad \frac{\partial}{\partial v_j}\ell(u,v) \;=\; n_{+,j}\frac{1}{v_j}$$

Führen wir zwei Lagrange-Multiplikatoren μ_1 und μ_2 ein um die Nebenbedingungen $u_1 + \ldots u_m = 1$ und $v_1 + \ldots v_n = 1$ zu berücksichtigen, so finden wir das Gleichungssystem

$$\mu_1 \;=\; \frac{\partial}{\partial u_i}\ell(u,v) \;=\; n_{i,+}\frac{1}{u_i}$$

$$\mu_2 \;=\; \frac{\partial}{\partial v_j}\ell(u,v) \;=\; n_{+,j}\frac{1}{v_j}$$

Zu fixiertem $r \in 1 \ldots m$ also $n_{r,+}/u_r = \mu_1 = n_{i,+}/u_i$ und damit $n_{i,+} = n_{r,+} \cdot u_i/u_r$. Durch Summation wird daraus

$$N \;=\; \sum_{i=1}^{m} n_{i,+} \;=\; \sum_{i=1}^{m} n_{r,+}\frac{u_i}{u_r} \;=\; n_{r,+}\frac{1}{u_r}\sum_{i=1}^{m} u_i \;=\; n_{r,+}\frac{1}{u_r}$$

und damit $u_r = n_{r,+}/N$. Genauso findet man $v_s = n_{+,s}/N$ zu fixiertem $s \in 1 \ldots n$. Da dies der einzige kritische Punkt von L ist, und L am Rand verschwindet, muss es sich dabei um ein Maximum handeln.

□

Die obige Likelihood Funktion $L(u,v)$ ist (bis auf einen Vorfaktor) die Eintrittswahrscheinlichkeit, dass unter den N Versuchen $n_{i,+}$ mal das Ereignis A_i und $n_{+,j}$ mal das Ereignis B_j eintritt - unter der Annahme, dass die A_i und B_j unabhängig sind. In so fern macht es Sinn diese Wahrscheinlichkeit zu maximieren. Zur Regression haben wir aber einen anderen Ansatz verwendet: Wir können auch den quadratischen Abstand der Schätzer $u_i \cdot v_j$ von den realen Häufigkeiten $h_{i,j}$ minimieren. Es stellt sich heraus, dass die Randhäufigkeiten, dafür *nicht* die besten Schätzer darstellen. Statt dessen ergibt sich ein Eigenwertproblem, wie der folgende Satz zeigt:

Satz 16.4:
Sei $H = (h_{i,j}) \in \mathrm{mat}_m n(\mathbb{R})$ eine Kontingenztabelle, wobei $1 \leq m,\, n \in \mathbb{N}$ beliebig sind. D.h. für alle $i \in 1 \ldots m$ und $j \in 1 \ldots n$ gelte $0 \leq h_{i,j} \leq 1$ und

$$\sum_{i=1}^{m}\sum_{j=1}^{n} h_{i,j} \;=\; 1$$

Wir nehmen an H ist die unabhängige Kombination zweier Merkmale, d.h. wir suchen $u = (u_1, u_2, \ldots, u_m) \in \mathbb{R}^m$ und $v = (v_1, v_2, \ldots, v_n) \in \mathbb{R}^n$ so dass für alle $i \in 1 \ldots m$ und $j \in 1 \ldots n$ gilt $h_{i,j} \approx u_i v_j$. Dazu betrachten wir die quadratische Abweichung von der unabhängigen Kombination

$$q(u,v) \;:=\; \sum_{i=1}^{m} \sum_{j=1}^{n} (h_{i,j} - u_i v_j)^2$$

Diese hat ein (nicht notwendig eindeutiges) globales Minimum $(u,v) \in \mathbb{R}^{m+n}$, das wir als *MLE-Schätzer* bezeichnen. Dabei sind folgende Aussagen für $u \neq 0$ äquivalent

(a) (u,v) ist ein kritischer Punkt von q

(b) $Hv \;= \|v\|^2\, u$ und $H^* u \;= \|u\|^2\, v$

(c) u ist ein Eigenvektor von HH^* und $v \;= \|u\|^{-2}\, H^* u$

Und ist $u \neq 0$ ein Eigenvektor von HH^* zum Eigenwert λ und ist v definiert, durch $v = \|u\|^{-2}\, H^* u$, dann gilt für die quadratische Abweichung

$$q(u,v) \;=\; \|H\|_2^2 - \lambda$$

Und damit erhalten wir unmittelbar eine Lösungsstrategie zur Bestimmung des MLE-Schätzers der unabhängigen Kombinationen u und v in H:

1. Berechne die Eigenwerte von HH^*
2. Wähle den größten Eigenwert λ von HH^*
3. Berechne einen Eigenvektor e von HH^* zum Eigenwert λ
4. setze $u \;:= \|e\|_1^{-1}\, e$
5. setze $v \;:= \|u\|^{-2}\, H^* u$

Beispiel 16.5:
Das folgende Beispiel zeigt, dass die *Randhäufigkeiten* nicht der beste Schätzer für die unabhängige Kombination einer Kontingenztabelle sind. Und die Abweichung tritt bereits bei einer *Vierfeldertafel* auf:

0.1	0.2	0.3
0.3	0.4	0.7
0.4	0.6	1.0

Die Randhäufigkeiten sind also $h_{\bullet,+} = (0.3, 0.7)$ und $h_{+,\bullet} = (0.4, 0.6)$ und man berechnet $q(h_{\bullet,+}, h_{+,\bullet}) \approx 0.0016$ (auf 4 Stellen gerundet). Betrachten wir also die positive, symmetrische Matrix HH^*

$$HH^* \;=\; \begin{pmatrix} 0.1 & 0.2 \\ 0.3 & 0.4 \end{pmatrix} \begin{pmatrix} 0.1 & 0.3 \\ 0.2 & 0.4 \end{pmatrix} \;=\; \begin{pmatrix} 0.05 & 0.11 \\ 0.11 & 0.25 \end{pmatrix}$$

Diese hat die Eigenwerte $\mathrm{Spec}(HH^*) = \{\, 0.013, 0.2987 \,\}$ (wieder auf 4 Stellen gerundet). Der größere der beiden Werte ist $\lambda = 0.2987$, zu diesem berechnen wir einen Eigenvektor $e = (0.11, 0.2487)$. Aus diesem bilden wir $u = (0.3067, 0.6933)$ und verarbeiten diesen weiter zu $v = (0.4153, 0.5892)$. Man beachte, das $\|u\|_1 = 1$ ist (nach Konstruktion), aber $\|v\|_1 > 1$. Das ist der verbleibende Haken an diesem Ansatz. Man sieht aber

$$q(u,v) \;\approx\; 0.0013 \;<\; 0.0016 \;\approx\; q(h_{\bullet,+}, h_{+,\bullet})$$

Und auch wenn wir $w := \| v \|_1^{-1}\, v$ setzen (so dass sich v wieder als Wahrscheinlichkeitsverteilung interpretieren lässt), so ist $q(u,w) < q(h_{\bullet,+}, h_{+,\bullet})$. Die Randhäufigkeiten liefern also auch dann nicht die beste Schätzung für $h_{i,j} \approx u_i v_j$, wenn wir zusätzlich $u \in \Delta_m$ und $v \in \Delta_n$ fordern.

Beweis:
Wir bezeichnen $r^2 := \| u \|^2 + \| v \|^2$, den (euklidschen) Abstand von (u,v) von 0 zum Quadrat. Ferner seien $\mu := u/r$ und $\nu := v/r$. Dann kann man $q(u,v)$ umschreiben, zu

$$\begin{aligned}
q(u,v) &= \sum_{i=1}^{m}\sum_{j=1}^{n}(h_{i,j} - u_i v_j)^2 \\
&= \sum_{i=1}^{m}\sum_{j=1}^{n}(h_{i,j}^2 + u_i^2 v_j^2 - 2h_{i,j}u_i v_j) \\
&= \|H\|^2 + \sum_{i=1}^{m}\sum_{j=1}^{n} u_i v_j (u_i v_j - 2h_{i,j}) \\
&= \|H\|^2 + r^4 \sum_{i=1}^{m}\sum_{j=1}^{n} \mu_i \nu_j \left(\mu_i \nu_j - \frac{2h_{i,j}}{r^2}\right)
\end{aligned}$$

Nach Konstruktion sind μ_i und $\nu_j \in [0,1]$. Damit ist die verbleibende Doppelsumme (als stetige Funktion auf einem Kompaktum) beschränkt. Und für $r \to \infty$ konvergiert sie stets gegen $\| \mu \|^2 \| \nu \|^2$. Also besteht für $r \to \infty$ die asymptotische Gleichheit

$$q(r\mu, r\nu) \;\sim\; \|H\|^2 + r^4\, \|\mu\|^2 \|\nu\|^2$$

Ist $\mu = 0$ oder $\nu = 0$, dann folgt $q(u,v) = \| H \|^2$. Und andernfalls wächst $q(u,v)$ für $r \to \infty$ aber unbeschränkt. Insbesondere konvergiert $q(u,v)$ für $\| u \| \to \infty$ und $\| v \| \to \infty$ niemals gegen 0. Nach unten ist aber durch $q(u,v) \geq 0$ beschränkt. Insgesamt muss q ein globales Minimum annehmen.

Wir zeigen nun die Äquivalenz der Aussagen, beginnend mit (a) $\iff$ (b): Ein kritischer Punkt liegt vor, wenn der Gradient $\nabla q(u,v) = 0$ verschwindet, d.h. wenn alle partiellen Ableitungen an dieser Stelle 0 sind. Diese lauten

$$0 \;=\; \partial_{u_r} q(u,v) \;=\; \sum_{j=1}^{n}(2u_r v_j^2 - 2h_{r,j}v_j)$$

$$0 \;=\; \partial_{v_s} q(u,v) \;\;=\;\; \sum_{i=1}^{m} (2u_i^2 v_s - 2h_{i,s}u_i)$$

Die erste Gleichung lässt sich aber leicht äquivalent umformen, zur r-ten Komponente der Vektorgleichung $Hv = \|v\|^2\, u$. Die zweite Gleichung liefert auf dieselbe Weise die s-te Komponente von $H^*u = \|u\|^2\, v$.

$$\|v\|^2\, u_r \;=\; \sum_{j=1}^{n} u_r v_j^2 \;=\; \sum_{j=1}^{n} h_{r,j} v_j \;=\; [Hv]_r$$

$$\|u\|^2\, v_s \;=\; \sum_{i=1}^{m} u_i^2 v_s \;=\; \sum_{i=1}^{m} h_{i,s} u_i \;=\; [H^*u]_s$$

Für (b) $\Longrightarrow$ (c) ist nur zu zeigen, dass u ein Eigenvektor von HH^* ist. Dies ist aber denkbar einfach - wir rechnen auch gleich nach, dass v ein Eigenvektor zum selben Eigenwert $\lambda := \|u\|^2\|v\|^2$ ist:

$$HH^*u \;=\; H\,\|u\|^2\, v \;= \|u\|^2\, Hv \;= \|u\|^2\|v\|^2\, u \;=\; \lambda u$$

$$H^*Hv \;=\; H^*\,\|v\|^2\, u \;= \|v\|^2\, H^*u \;= \|v\|^2\|u\|^2\, v \;=\; \lambda v$$

Wir schließen nun den Kreis mit (c) $\Longrightarrow$ (b): Sei also $HH^*u = \lambda u$ und v definiert, durch $v = \|\, u\,\|^{-2}\; H^*u$. Wir zeigen zunächst, dass dann auch $H^*Hv = \lambda v$ ein Eigenvektor (oder $v = 0$) sein muss:

$$\begin{aligned} H^*Hv \;&=\; H^*H\frac{1}{\|u\|^2}H^*u \;=\; \frac{1}{\|u\|^2}H^*HH^*u \\ &=\; \frac{1}{\|u\|^2}H^*\lambda u \;=\; \lambda\frac{1}{\|u\|^2}H^*u \;=\; \lambda v \end{aligned}$$

Damit wird aber auch $\|u\|^2\|v\|^2 = \lambda$ automatisch erfüllt, denn man sieht

$$\begin{aligned} \|u\|^2\|v\|^2 \;&=\; \|u\|^2\,\frac{1}{\|u\|^4}\,\|H^*u\|^2 \\ &=\; \frac{1}{\|u\|^2}\langle H^*u \mid H^*u\rangle \\ &=\; \frac{1}{\|u\|^2}\langle HH^*u \mid u\rangle \\ &=\; \frac{1}{\|u\|^2}\langle \lambda u \mid u\rangle \\ &=\; \lambda\frac{1}{\|u\|^2}\langle u \mid u\rangle \;=\; \lambda \end{aligned}$$

Nach Annahme haben wir $H^*u \;=\; \|\, u\,\|^2\; v$ bereits, es bleibt nur noch die zweite Behauptung $Hv \;= \|v\|^2\, u$ zu zeigen:

$$Hv \;=\; H\frac{1}{\|u\|^2}H^*u \;=\; \frac{1}{\|u\|^2}HH^*u \;=\; \frac{1}{\|u\|^2}\lambda u \;= \|v\|^2\, u$$

Es fehlt nur noch die Identität für $q(u,v)$ wenn u ein Eigenvektor zum Eigenwert λ und v wie eben ist. Doch diese lässt sich direkt nachrechnen:

$$\begin{aligned}
q(u,v) &= \sum_{i=1}^{m}\sum_{j=1}^{n}(h_{i,j}-u_iv_j)^2 \\
&= \sum_{i=1}^{m}\sum_{j=1}^{n}\left(h_{i,j}^2-2h_{i,j}u_iv_j+u_i^2v_j^2\right) \\
&= \sum_{i=1}^{m}\sum_{j=1}^{n}h_{i,j}^2-2\sum_{i=1}^{m}\sum_{j=1}^{n}h_{i,j}u_iv_j+\sum_{i=1}^{m}\sum_{j=1}u_i^2v_j^2 \\
&= \|H\|_2^2-2\sum_{j=1}^{n}v_j\sum_{i=1}^{m}h_{i,j}u_i+\|u\|^2\|v\|^2 \\
&= \|H\|_2^2-2\langle v \mid H^*u\rangle+\lambda \\
&= \|H\|_2^2-2\,\|u\|^{-2}\,\langle H^*u \mid H^*u\rangle+\lambda \\
&= \|H\|_2^2-2\,\|u\|^{-2}\,\langle HH^*u \mid u\rangle+\lambda \\
&= \|H\|_2^2-2\,\|u\|^{-2}\,\langle \lambda u \mid u\rangle+\lambda \\
&= \|H\|_2^2-2\lambda\,\|u\|^{-2}\,\langle u \mid u\rangle+\lambda \\
&= \|H\|_2^2-2\lambda+\lambda\ =\|H\|_2^2-\lambda
\end{aligned}$$

□

Test 16.6: χ^2-Test auf Unabhängigkeit
Wir betrachten zwei Zufallsvariablen $X : R \to S$ und $Y : R \to T$ auf demselben (wenn auch beliebigem) Wahrscheinlichkeitsraum (R, Σ_R, p). Die Frage ist nun, ob X und Y stochastisch unabhängig sind. Natürlich lässt sich das ohne weiteres nicht mit Sicherheit sagen, aber wir können wieder einen Hypothesentest machen. Es sei also eine Stichprobe $(x_k \mid y_k) \in S \times T$ vorgelegt, wobei $k \in 1\ldots n$ und $1 \leq n \in \mathbb{N}$. Zudem wählen wir endliche Partitionen von S und T, diese seien

$$S = \bigcup_{i=1}^{s} C_i \quad \text{bzw.} \quad T = \bigcup_{j=1}^{t} D_j$$

und wir bezeichnen die tatsächliche Häufigkeit der Stichprobe in der Klasse $C_i \times D_j$ mit

$$n_{i,j} := \#\{\, k \in 1\ldots n \mid x_k \in C_i \text{ und } y_k \in D_j \,\}$$

und die relative Häufigkeit, mit $p_{i,j} := n_{i,j}/n$. Wenn X und Y unabhängig wären, dann müsste $n_{i,j}$ ungefähr das Produkt der Randhäufigkeiten sein. Wir erwarten für $n_{i,j}$ also in etwa den Wert

$$e_{i,j} := \frac{n_{i,+}\cdot n_{+,j}}{n} = \frac{1}{n}\left(\sum_{v=1}^{t} n_{i,v}\right)\left(\sum_{u=1}^{s} n_{u,j}\right)$$

Genau genommen führen wir nun also einen Test für die folgende Hypothese durch: Für alle $i \in 1 \dots s$ und alle $j \in 1 \dots t$ sind die Ereignisse $X \in C_i$ und $Y \in D_j$ stochastisch unabhängig.
Dazu betrachten wir wieder die Pearson-Statistik der relativen quadratischen Abweichung der tatsächlichen von den erwarteten Häufigkeiten, in Formeln

$$\Delta \ := \ \sum_{i=1}^{s} \sum_{j=1}^{t} \frac{(n_{i,j} - e_{i,j})^2}{e_{i,j}}$$

Dabei ist die Zufallsvariable Δ wieder annähernd χ^2-verteilt, diesmal jedoch mit $f := (s-1)(t-1)$ Freiheitsgraden. Als Faustregel sagt man, die Annäherung ist gut genug, wenn gilt

$$\forall\, i \in 1 \dots s \; \forall\, j \in 1 \dots t \ : \ e_{i,j} \geq 5$$

Der Test funktioniert nun also wie folgt: Man legt ein Sicherheitsniveau $\alpha \in [0,1]$ fest und führt die Stichprobe $(x_1, y_1), \dots, (n_n, y_n) \in S \times T$ aus. Dabei ist die Anzahl n so zu wählen, dass gilt

$$\forall\, i \in 1 \dots s \; \forall\, j \in 1 \dots t \ : \ n \cdot p(X \in C_i) \cdot p(Y \in D_j) \geq 5$$

Wir nehmen die Hypothese der Unabhängigkeit von X und Y an, wenn $\Delta \leq \chi_f(1-\alpha)$ ist. Dabei ist χ_f die Verteilungsfunktion der χ^2-Verteilung zu $f := (s-1)(t-1)$ Freiheitsgraden, wie in (13.9) beschrieben. Ist hingegen $\Delta > \chi_f(1-\alpha)$, dann ist die beobachtete Abweichung Δ (unter der Annahme der Unabhängigkeit) unwahrscheinlicher, als α, so dass X und Y als anhängig angesehen werden müssen.

Beispiel 16.7: Vier-Felder-Tafel
Im einfachsten Fall wurden die Ereignisräume S und T in jeweils zwei Klassen unterteilt: $S = C_1 \cup C_2$ und $T = D_1 \cup D_2$. Dann können die tatsächlichen Häufigkeiten der Stichprobe in einer 2×2-Matrix dargestellt werden:

	$Y \in D_1$	$Y \in D_2$
$X \in C_1$	a	b
$X \in C_2$	c	d

dabei ist dann natürlich $n = a + b + c + d$ die Anzahl der Stichproben insgesamt. Und die bei Unabhängigkeit von X und Y erwarteten Häufigkeiten sind gegeben, durch

$$\begin{aligned} e_a &= \frac{(a+b)(a+c)}{n} \\ e_b &= \frac{(a+b)(b+d)}{n} \\ e_c &= \frac{(a+c)(c+d)}{n} \\ e_d &= \frac{(b+d)(c+d)}{n} \end{aligned}$$

$$\Delta \ = \ \frac{(a-e_a)^2}{e_a} + \frac{(b-e_b)^2}{e_b} + \frac{(c-e_c)^2}{e_c} + \frac{(d-e_d)^2}{e_d}$$

Dabei ist beispielsweise $(a-e_a)^2/e_a = (a^2 - 2ae_a + e_a^2)/e_a = a^2/e_a - 2a + e_a$ und damit $\Delta = (a^2/e_a + b^2/e_b + c^2/e_c + d^2/e_d) - 2n + (e_a + e_b + e_c + e_d)$. Wegen $(a+b)(a+c)+(a+b)(b+d)+(a+c)(c+d)+(b+d)(c+d) = n^2$ wird daraus $\Delta = (a^2/e_a + b^2/e_b + c^2/e_c + d^2/e_d) - n$. Und eine explizite Rechnung liefert daraus dann die einfachere Formel

$$\Delta \;=\; n \cdot \frac{(ad-bc)^2}{(a+b)(a+c)(b+d)(c+d)}$$

In diesem Fall haben wir $f = (2-1)\cdot(2-1) = 1$ Freiheitsgrad, legen wir das Sicherheitsviveau α fest, dann muss die Hypothese der Unabhängigkeit von X und Y (bei dieser Klassierung) *verworfen* werden, wenn

$$\chi_1(1-\alpha) \;<\; \Delta \;=\; n \cdot \frac{(ad-bc)^2}{(a+b)(a+c)(b+d)(c+d)}$$

Test 16.8: Exakter Fisher Test:
Der Chi-Quadrat-Test auf Unabhängigkeit verwendet die Ähnlichkeit der Verteilung der Test-Statistik zu einer Chi-Quadrat-Verteilung, die gegeben ist, solange die Bedingungen $e_{i,j} \geq 5$ gegeben sind. Selbst dann ist der Test aber nicht exakt, α keine obere Schranke der Irrtumswahrscheinlichkeit sondern nur eine Abschätzung. Abhilfe schafft der folgende Test:

Wie im Chi-Quadrat-Test auf Unabhängigkeit haben wir $X : R \to S$ und $Y : R \to T$ und Partitionen (C_i) von S bzw. (D_j) von T. Die Kontingenztabelle habe wieder die Einträge $n_{i,j}$ wobei $i \in 1 \ldots r$ und $j \in 1 \ldots s$

	$Y \in D_1$	$\ldots$	$Y \in D_j$	$\ldots$	$Y \in D_s$	
$X \in C_1$	$n_{i,1}$	$\ldots$		$\ldots$	$n_{1,s}$	u_1
$\vdots$	$\vdots$		$\vdots$		$\vdots$	
$X \in C_i$		$\ldots$	$n_{i,j}$	$\ldots$		u_i
$\vdots$	$\vdots$		$\vdots$		$\vdots$	
$X \in C_r$	$n_{r,1}$	$\ldots$		$\ldots$	$n_{r,s}$	u_r
	v_1		v_j		v_s	n

Wir haben die Randhäufigkeiten also mit $u_i = n_{i,+}$ bzw. $v_j = n_{+,j}$ bezeichnet und setzten $u = (u_1, \ldots, u_m)$ und $v = (v_1, \ldots, v_n)$. Wir betrachten nun also den Zustandsraum $\mathcal{K}$ aller Kontingenztafeln dieser Größe und sortieren diejenigen aus, die dieselben Randhäufigkeiten haben:

$$\begin{aligned} \mathcal{K} &:= \{\, A \in \mathrm{mat}_{r,}(\mathbb{Z}) \mid a_{i,j} \geq 0 \,\} \\ \mathcal{K}_{u,v} &:= \{\, A \in \mathcal{K} \mid a_{i,+} = u_i,\ a_{+,j} = v_j \,\} \end{aligned}$$

Die Idee ist nun Folgende: Wir nehmen an H_0, die Zufallsvariablen X und Y seien unabhängig voneinander und fixieren die Randwerte u und v. Dann ist die Wahrscheinlichkeit die Matrix $A = (a_{i,j}) \in \mathcal{K}_{u,v}$ zu beobachten gegeben, durch (vgl. *mathworld.wolfram.com/FishersExactTest.html*)

$$p(A) \;=\; \frac{\prod_{i=1}^r u_i! \cdot \prod_{j=1}^s v_j!}{n! \cdot \prod_{i=1}^r \prod_{j=1}^s a_{i,j}!}$$

Wir fragen uns also, ob die beobachtete Matrix $N = (n_{i,j}) \in \mathcal{K}_{u,v}$ besonders ungewöhnlich ist, oder nicht. D.h. wir summieren die Wahrscheinlichkeiten aller Matrizen A mit (denselben Randwerten), die höchstens so wahrscheinlich sind wie N

$$\delta \;:=\; \sum_{A \in \mathcal{K}_{u,v}} \begin{cases} p(A) & \text{für } p(A) \leq p(N) \\ 0 & \text{für } p(A) > p(N) \end{cases}$$

Ist α unsere vorgelegtes Sicherheitsniveau und ist $\delta \geq \alpha$, dann gibt es keinen Grund an H_0 zu zweifeln. Ist hingegen $\delta < \alpha$, dann müssen wir H_0 verwerfen, X und Y sind abhängig.

Bemerkung 16.9:
Speziell im Fall $r = s = 2$ einer 2×2 Kontingenztafel lässt sich der Raum $\mathcal{K}_{u,v}$ der Matrizen leicht durchnummerieren, es genügt:

$$\begin{aligned} a_{1,1} &\in \max\{\, 0, u_1 - v_2 \,\} \ldots \min\{\, u_1, v_1 \,\} \\ a_{1,2} &= u_1 - a_{1,1} \\ a_{2,1} &= v_1 - a_{1,1} \\ a_{2,2} &= v_2 - a_{1,2} \end{aligned}$$

zu betrachten, das sind $1 + \min\{\, u_1, u_2, v_1, v_2 \,\}$ Matritzen. Darunter ist natürlich auch die Matrix N der Beobachtungswerte. Zu diesen berechnet man jeweils $p(A)$ und summiert alle $p(A)$ für die $p(A) \leq p(N)$ ist. Die dabei entstehende Summe ist δ.

Kapitel 17

Integrationstheorie

In diesem Kapitel wollen wir ein wenig den Begriff eines Integrals betrachten, der ja notwendig wird, um den Begriff der Summe zu verallgemeinern, auf den Fall dass die Indexmenge der Summanden kontinuierlich (überabzählbar) ist. Dazu betrachten wir zunächst den *quick-and-dirty* Ansatz von Riemann, den wohl bekanntesten Ansatz vermöge Regelfunktionen und schließlich den raffiniertesten Ansatz von Lebesgue.

Definition 17.1:

(i) Sind $a = (a_1, \ldots, a_d)$ und $b = (b_1, \dot{;} b_d) \in \mathbb{R}^d$ vorgelegte d-Tupel, mit $a \leq b$ (d.h. für alle $i \in 1 \ldots d$ gilt $a_i \leq b_i$). Dann definieren wir den (kompakten) **Quader** $[a, b] \subseteq \mathbb{R}^d$ mit den Eckpunkten a und b als

$$[a, b] \;:=\; \Big\{ x = (x_1, \ldots, x_d) \in \mathbb{R}^d \mid \forall\, i \in 1 \ldots d : a_i \leq x_i \leq b_i \Big\}$$

Und die Menge aller kompakten Quader in $\mathbb{R}^d$ bezeichnen wir mit $\mathcal{Q}^d := \big\{ [a, b] \mid a \leq b \in \mathbb{R}^d \big\}$. Das **Volumen** eines Quaders ist das Produkt seiner Kantenlängen, d.h. wir definieren weiter

$$\mu([a, b]) \;:=\; \prod_{i=1}^{d} (b_i - a_i)$$

(ii) Ist nun $A \subseteq \mathbb{R}^d$ eine beliebige Teilmenge von $\mathbb{R}^d$, dann definieren wir deren **charakteristische Funktion** als

$$\mathbf{1}_A \;:\; \mathbb{R}^d \to \mathbb{R} \;:\; x \mapsto \begin{cases} 1 & \text{für } x \in A \\ 0 & \text{für } x \notin A \end{cases}$$

(iii) Wir nennen eine Funktion $t : \mathbb{R}^d \to \mathbb{R}$ eine **Treppenfunktion**, falls es endlich viele Quader $Q(1), \ldots, Q(m) \in \mathcal{Q}^d$ und Zahlen $a_1, \ldots, a_m \in \mathbb{R}$ gibt, so dass sich t schreiben lässt, als

$$t \;=\; \sum_{k=1}^{m} a_k \mathbf{1}_{Q(k)}$$

Die Menge aller Treppenfunktionen wird also erzeugt als der von den charakteristischen Funktionen von Quadern aufgespannte Vektorraum. Wir bezeichnen sie mit

$$\mathcal{T}(\mathbb{R}^d, \mathbb{R}) \;:=\; \mathrm{Lh}\left\{ \mathbf{1}_Q \mid Q \in \mathcal{Q}^d \right\}$$

(iv) Und ist $t \in \mathcal{T}(\mathbb{R}^d, \mathbb{R})$ eine Treppenfunktion, dann gibt es eine Darstellung $t = \sum_k a_k \mathbf{1}_{Q(k)}$ wie oben, so dass für alle $j, k \in 1 \ldots m$ gilt $a_k \neq 0$ und $j \neq k \implies \mu(Q(j) \cap Q(k)) = 0$. Und für diese definieren wir dann deren **Integral** durch

$$\int t \;:=\; \sum_{k=1}^{m} a_k \cdot \mu(Q(k)) \;\in\; \mathbb{R}$$

(v) **Riemann-Integral:** Sind nun f und g zwei beliebige Funktionen, dann schreiben wir $f \leq g$ falls für alle $x \in \mathbb{R}^d$ gilt $f(x) \leq g(x)$. Und für $f : \mathbb{R}^d \to \mathbb{R}$ definieren wir die folgenden beiden Grenzwerte

$$\begin{aligned} I^*(f) \;&:=\; \sup\left\{ \int t \;\middle|\; t \in \mathcal{T}(\mathbb{R}^d, \mathbb{R}) \text{ mit } t \leq f \right\} \\ I_*(f) \;&:=\; \inf\left\{ \int t \;\middle|\; t \in \mathcal{T}(\mathbb{R}^d, \mathbb{R}) \text{ mit } f \leq t \right\} \end{aligned}$$

Wir nennen f **Riemann-integrierbar**, falls $I^*(f) = I_*(f)$ ist. In diesem Fall definieren wir dann das **Riemann-Integral**, durch

$$\int f \;:=\; I^*(f) \;=\; I_*(f)$$

(vi) **Regel-Integral:** Ist $f : \mathbb{R}^d \to \mathbb{R}$ eine beliebige Funktionen, dann setzen wir $\| f \|_\infty := \sup\{ |f(x)| \mid x \in \mathbb{R}^d \}$. Und wir sagen die Folge (f_n) konvergiert *gleichmäßig* gegen f, falls gilt

$$(f_n) \to f \quad :\Longleftrightarrow \quad \lim_{n\to\infty} \| f - f_n \|_\infty \;=\; 0$$

Wir nennen f nun **(Regel-)integrierbar**, falls es eine Folge $(t_n) \subseteq \mathcal{T}(\mathbb{R}^d, \mathbb{R})$ von Treppenfunktionen gibt, die gleichmäßig $(t_n) \to f$ gegen f konvergiert. Dann ist der Grenzwert unabhängig von der speziellen Folge und wir nennen ihn das **(Regel-)Integral** von f:

$$\int f \;:=\; \lim_{n\to\infty} \int t_n$$

Bemerkung 17.2:
Wir haben also zwei Integralbegriffe eingeführt, für Funktionen $f : \mathbb{R}^d \to \mathbb{R}$. Üblich ist aber auch folgende Schreibweise, in der die Variable der Funktion betont wird

$$\int f(x)\, dx \;:=\; \int f$$

Von diesen beiden Begriffen ist das Regel-Integral die zumeist verwendete Definition. Dieses Konzept wird beispielsweise verfolgt, in [Fischer, Kaul, Mathematik für Physiker, Band 1, Kapitel 23]. Ist f aber nur auf einer Teilmenge $f : D \to \mathbb{R}$ mit $D \subseteq \mathbb{R}^d$ definiert, dann setzen wir f auf ganz $\mathbb{R}^d$ durch 0 fort:

$$\overline{f}(x) \quad := \quad \begin{cases} f(x) & \text{für } x \in D \\ 0 & \text{für } x \notin D \end{cases}$$

Dann heißt f **integrierbar**, wenn ihre Fortsetzung $\overline{f}$ integrierbar ist und in diesem Fall kann auch das **Integral** von f definiert werden, als

$$\int_D f \quad := \quad \int \overline{f}$$

Definition 17.3:
Sei $(S, \mathcal{A}, \mu)$ ein Raum mit Inhalt. Eine Funktion $e : S \to \mathbb{R}$ heißt **einfach**, wenn sie eine endliche Bildmenge $e(S) \subseteq \mathbb{R}$ besitzt und zu jedem $a \in \mathbb{R}$ ist die Faser $e^{-1}(a) \in \mathcal{A}$ messbar. Die Menge aller einfachen Funktionen bezeichnen wir, mit

$$\mathcal{E}(S, \mathbb{R}) \quad := \quad \{\, e \in \mathcal{M}(S, \mathbb{R}) \mid \#e(S) < \infty \,\}$$

In diesem Fall können wir das **Integral** einer einfachen Funktion $e : S \to \mathbb{R}$ erklären, als die folgende (endliche) Summe:

$$\int e \, d\mu \quad := \quad \sum_{a \in e(S)} a \cdot \mu(e^{-1}(a))$$

Beispiel 17.4:
Ist $S = \mathbb{R}^d$ der d-dimensionale Raum und ist $\mathcal{A} = \mathcal{B}(\mathbb{R}^d)$ die Borel-σ-Algebra auf $\mathbb{R}^d$ unter dem Lebesgue-Maß $\mu = \lambda^d$, dann sind Treppenfunktionen insbesondere auch einfache Funktionen:

$$\mathcal{T}(\mathbb{R}^d, \mathbb{R}) \quad \subseteq \quad \mathcal{E}(\mathbb{R}^d, \mathbb{R})$$

und das Integral einer Treppenfunktion t stimmt mit dem Integral als einfache Funktion überein. Anschaulich gesprochen kann man sich die Situation wie folgt vorstellen: Bei Treppenfunktionen wurde der Definitionsbereich $\mathbb{R}^d$ in Quader zerlegt. Treppenfunktionen sind also schön blockweise. Bei einfachen Funktionen wurde der Wertebereich $\mathbb{R}$ zerlegt - die Fasern müssen keine Quader sein, sondern können eine kompliziertere Gestalt haben. Indem man die Komplikation aber in zwei Teile zerlegt (erst das Maß auf der Definitionsmenge, dann das Integral) kann man diesen Integralbegriff aber besser kontrollieren und er wird bessere Eigenschaften haben.

Bemerkung 17.5:
Im Grunde könnte man noch ein wenig allgemeiner starten: Für den Begriff einer einfachen Funktion und deren Integral muss das Tripel $(S, \mathcal{A}, \mu)$ lediglich die folgenden drei Eigenschaften erfüllen:

(1) $S \neq \emptyset$ ist eine nicht-leere Menge.

(2) $\mathcal{A} \subseteq \mathcal{P}(S)$ ist eine Familie von Teilmengen von S für die gilt $\emptyset \in \mathcal{A}$ und sind A, $B \in \mathcal{A}$, dann sind auch $A \cup B$ und $A \cap B \in \mathcal{A}$.

(3) μ ist eine Abbildung der Form $\mu : \mathcal{A} \to R$, wobei R ein kommutativer Ring ist. Dabei ist $\mu(\emptyset) = 0$ und für alle A, $B \in \mathcal{A}$ mit $A \cap B = \emptyset$ gilt $\mu(A \cup B) = \mu(A) + \mu(B)$.

In diesem Fall können wir einfache Funktionen $f : S \to R$ mit Werten in einem kommutativen Ring R betrachten und auch das Integral wird einen Wert in R annehmen. In dieser Form wird beispielsweise das motivische Integral auf Varietäten gebildet.

Definition 17.6:
Sei $(S, \mathcal{A}, \mu)$ ein Maßraum, dann nennen wir eine Abbildung ω der Form $\omega : S \to \mathbb{R}^+ \cup \{\infty\}$ eine *numerische* Funktion auf S. Ist ω sogar $\mathcal{A}$-messbar (d.h. für alle a, $b \in \mathbb{R}^+ \cup \{\infty\}$ gilt $\omega^{-1}([a,b]) \in \mathcal{A}$), dann nennen wir ω eine **Belegungsfunktion** von S. Die Menge aller Belegungsfunktionen von S bezeichnen wir mit

$$\Omega(S) \quad := \quad \left\{ \omega : S \to \mathbb{R}^+ \cup \{\infty\} \mid \forall\, 0 \leq a \leq b \leq \infty : \omega^{-1}([a,b]) \in \mathcal{A} \right\}$$

Ist nun $\omega \in \Omega(S)$ eine Belegungsfunktion, dann gibt es nach [Bauer, Maßtheorie, 11.6] eine Folge $(e_n) \subseteq \mathcal{E}(S, \mathbb{R})$ einfacher Funktionen, so dass gilt

(1) Die e_n sind positiv und bilden eine isoton steigende Folge, d.h. es gilt:

$$0 \;\leq\; e_0 \;\leq\; e_1 \;\leq\; \ldots \;\leq\; e_n \;\leq\; e_{n+1} \;\leq\; \ldots$$

(2) Die e_n konvergieren punktweise von unten gegen ω, oder anders gesagt: ω ist das Supremum der e_n. D.h. für alle $x \in S$ gilt die Gleichung

$$\omega(x) \quad = \quad \sup\{\, e_n(x) \mid n \in \mathbb{N} \,\}$$

Und in diesem Fall definieren wir das **Oberintegral** (nach [Bauer, Maßtheorie, 11.2] ist dies unabhängig von der speziell gewählten Folge) von ω

$$I^*(\omega) \quad := \quad \sup\left\{ \int e_n d\mu \mid n \in \mathbb{N} \right\} \quad \in \quad \mathbb{R} \cup \{\infty\}$$

Ist nun $f : S \to \mathbb{R}$ eine beliebige messbare Funktion auf S, dann sind nach [Bauer, Maßtheorie, 9.6] auch deren Positiv- und Negativteil messbar

$$f_+ : S \to \mathbb{R}^+ : x \mapsto \begin{cases} f(x) & \text{für } f(x) \geq 0 \\ 0 & \text{für } f(x) < 0 \end{cases}$$

$$f_- : S \to \mathbb{R}^+ : x \mapsto \begin{cases} 0 & \text{für } f(x) \geq 0 \\ -f(x) & \text{für } f(x) < 0 \end{cases}$$

Und wir nennen f **(Lebesgue-)integrierbar**, wenn f messbar ist, und die Oberintegrale der Positiv- und Negativteile endlich sind, d.h. falls $I^*(f_+)$ und $I^*(f_-) \in \mathbb{R}$ sind. In diesem Fall definieren wir das **(Lebesgue-)Integral**

$$\int f \, d\mu \;\; := \;\; I^*(f_+) - I^*(f_-)$$

Die Menge der Lebesgue-integierbaren Funktionen auf S bezeichnen wir mit

$$\mathcal{L}(S, \mathbb{R}) \;\; := \;\; \{ f \in \mathcal{M}(S, \mathbb{R}) \mid I^*(f_+), I^*(f_-) \in \mathbb{R} \}$$

Bemerkung 17.7:
Auch in diesem Fall müssen wir die Notation noch ein wenig erweitern: Gelegentlich ist es notwendig, das Maß zu fixieren, auf das sich die Integrierbarkeit bezieht. In diesem Fall schreiben wir auch $I^*_\mu(\omega)$ anstelle $I^*(\omega)$ und $\mathcal{L}_\mu(S, \mathbb{R})$ anstelle $\mathcal{L}(S, \mathbb{R})$.

Ist nun $B \in \mathcal{A}$ eine messbare Menge dann ist die Restriktion von $\mathcal{A}$ auf B gerade $\mathcal{A} \cap B = \{ A \in \mathcal{A} \mid A \subseteq B \}$. Und wir erhalten wieder ein Maß μ_B auf B durch Restriktion von μ auf $\mathcal{A} \cap B$, d.h. $\mu_B : \mathcal{A} \cap B \to [0, \infty] : A \mapsto \mu(A)$. Damit ist auch das Integral über B definiert. Offensichtlich lässt sich dieses Integral aber auch über ganz S auswerten: Ist $f : B \to \mathbb{R}$, dann setzen wir f außerhalb von B durch 0 auf S fort:

$$\overline{f}(x) \;\; := \;\; \begin{cases} f(x) & \text{für } x \in B \\ 0 & \text{für } x \notin B \end{cases}$$

Dann ist $f \in \mathcal{L}(B, \mathbb{R})$, wenn ihre Fortsetzung $\overline{f} \in \mathcal{L}(S, \mathbb{R})$ ist und in diesem Fall stimmen dann auch die Integrale über B bzw. S überein

$$\int_B f \, d\mu_B \;\; = \;\; \int_S \overline{f} \, d\mu$$

Da keine Verwechslung möglich ist, werden wir an dieser Stelle den Index B beim restringierten Maß μ_B weglassen. D.h. wir bezeichnen μ_B wieder mit μ, obwohl das formal nicht ganz korrekt ist.

Es ist oftmals auch von Nutzen die Variable zu betonen, über die Integriert wird. In diesem Fall schreiben wir dann auch

$$\int_S f(x) \, d\mu(x) \;\; := \;\; \int f \, d\mu$$

Satz 17.8: [Bauer, MundI-Theorie, 12.3, 12.2, 13.2]
Sei $(S, \mathcal{A}, \mu)$ ein Maßraum und bezeichne $\mathcal{L}(S, \mathbb{R})$ die Menge der Lebesgue-integrierbaren Funktionen auf S. Dann gelten die folgenden Aussagen

(i) $\mathcal{L}(S, \mathbb{R})$ ist ein $\mathbb{R}$-Vektorraum und das Lebesgue-Integral ist eine isotone, lineare Abbildung von $\mathcal{L}(S, \mathbb{R})$ nach $\mathbb{R}$, für die gilt:

$$\int_B |f| \, d\mu \;\; \leq \;\; \mu(B) \cdot \| \, f \, \|_\infty$$

(ii) Ist $f : S \to \mathbb{R}$ eine beliebige Funktion, dann sind die folgenden fünf Aussagen äquivalent:

(a) $f \in \mathcal{L}(S, \mathbb{R})$ ist integrierbar
(b) f_+ und $f_- \in \mathcal{L}(S, \mathbb{R}^+)$ sind integrierbar
(c) es gibt positive, integrierbare Funktionen u und $v \in \mathcal{L}(S, \mathbb{R}^+)$ so dass f sich darstellen lässt, als: $f = u - v$
(d) $|f| \in \mathcal{L}(S, \mathbb{R}^+)$ ist integrierbar
(e) es gibt eine positive, integrierbare Funktion $p \in \mathcal{L}(S, \mathbb{R}^+)$ so dass f sich wie folgt abschätzen lässt: $|f| \leq p$

(iii) Ist $p \in \mathcal{L}(S, \mathbb{R}^+)$ eine positive, integrierbare Funktion, dann verschwindet das Integral über f genau dann, wenn sie fast überall 0 ist:

$$\int p \, d\mu \;=\; 0 \quad \Longleftrightarrow \quad \mu\left(p^{-1}(0)\right) \;=\; 0$$

Satz 17.9: [Bauer, MundI-Theorie, 16.4]
Sei $f : [a, b] \to \mathbb{R}$ eine Borel-messbare Funktion, auf dem kompakten Intervall $[a, b] \subseteq \mathbb{R}$. Ist f Riemann-integrierbar (zum Beispiel f stetig), dann ist f auch Lebesgue-integrierbar und die beiden Integrale haben den gleichen Wert. Insbesondere gilt für das Lebesgue-Integral wieder der *Fundamentalsatz der Analysis*: ist $f : [a, b] \to \mathbb{R}$ stetig und ist $F : [a, b] \to \mathbb{R}$ eine Stammfunktion von f (d.h. für alle $a \leq x \leq b$ gilt $F'(x) = f(x)$) dann gilt

$$\int_a^b f(x) \, d\lambda^1(x) \;\; = \;\; F(b) - F(a)$$

Definition 17.10:
Sei $(S, \mathcal{A}, \mu)$ ein Maßraum und $1 \leq p \in \mathbb{R}$ fixiert. Dann nennen wir eine Funktion $f : S \to \mathbb{R}$ **p-fach integrierbar**, falls $|f|^p$ integrierbar ist. Die Menge all dieser Funktionen bezeichnen wir mit

$$\mathcal{L}^p(S, \mathbb{R}) \;\; := \;\; \{ \, f \in \mathcal{M}(S, \mathbb{R}) \mid \; |f|^p \in \mathcal{L}(S, \mathbb{R}) \, \}$$

Insbesondere ist also $\mathcal{L}^1(S, \mathbb{R}) = \mathcal{L}(S, \mathbb{R})$. Bemerkung: ist $f : S \to \mathbb{R}$ messbar, dann ist $|f|^p$ bereits messbar. Es genügt also die Endlichkeit des Integrals zu testen. Ist nun $f \in \mathcal{L}^p(S, \mathbb{R})$, dann definieren wir die **p-Norm** von f durch

$$\|f\|_p \;:=\; \left(\int |f|^p \, d\mu \right)^{\frac{1}{p}}$$

Wir nennen $f : S \to \mathbb{R}$ **fast überall beschränkt**, falls es eine Zahl $m \in \mathbb{R}$ gibt, so dass f außerhalb einer Nullmenge durch m beschränkt ist:

$$\mu(\{\, x \in S \mid |f(x)| > m \,\}) \;=\; 0$$

Die Menge aller fast überall beschränkten Funktionen $f : S \to \mathbb{R}$ bezeichnen wir mit $\mathcal{L}^\infty(S, \mathbb{R})$. Und auf dieser Menge definieren wir

$$\|f\|_p \;:=\; \inf\{\, m \in \mathbb{R} \;\mid\; \mu(\{\, x \in S \mid |f(x)| > m \,\}) = 0 \,\}$$

Satz 17.11: [Bauer, MundI-Theorie, 14.1 bis 14.7]
Sei $(S, \mathcal{A}, \mu)$ ein Maßraum und $p \in [1, \infty]$, dann ist die Menge $\mathcal{L}^p(S, \mathbb{R})$ der p-fach integrierbaren Funktionen ein Untervektorraum von $\mathcal{F}(S, \mathbb{R})$. Und ist $\mu(S) < \infty$ endlich und sind $1 \le p \le q \le \infty$, dann gilt sogar

$$\mathcal{L}^q(S, \mathbb{R}) \;\subseteq\; \mathcal{L}^p(S, \mathbb{R}) \;\subseteq\; \mathcal{L}(S, \mathbb{R})$$

Sei μ wieder beliebig und $p \in [1, \infty]$, dann ist $\| \bullet \|_p$ eine *Halbnorm* auf $\mathcal{L}^p(S, \mathbb{R})$, d.h. für alle $a \in \mathbb{R}$ und alle Funktionen f, $g \in \mathcal{L}^p(S, \mathbb{R})$ gilt

$$\begin{aligned} \|af\|_p &= |a| \cdot \|f\|_p \\ \|f + g\|_p &\le \|f\|_p + \|g\|_p \end{aligned}$$

Vorsicht: es handelt sich dabei aber nicht um Normen, denn aus $\|f\|_p = 0$ folgt *nicht* $f = 0$, es folgt lediglich, dass $\{\, x \in S \mid f(x) \neq 0 \,\}$ eine μ-Nullmenge ist. Sind nun p und $q \in [1, \infty]$ so dass

$$\frac{1}{p} + \frac{1}{q} \;=\; 1$$

(wobei z.B. für $q = \infty$ noch $p = 1$ erlaubt ist) und sind $f \in \mathcal{L}^p(S, \mathbb{R})$ und $g \in \mathcal{L}^q(S, \mathbb{R})$ p-fach, bzw. q-fach integrierbar, dann ist das Produkt $fg \in \mathcal{L}(S, \mathbb{R})$ 1-fach integrierbar, denn es gilt die *Hölder'sche Ungleichung*

$$\|fg\|_1 \;\le\; \|f\|_p \cdot \|g\|_q$$

Satz 17.12: [Bauer, MundI-Theorie, 15.4 und 15.6]
Sei $(S, \mathcal{A}, \mu)$ ein Maßraum und $1 \le p \in \mathbb{R}$ fixiert. Wir betrachten eine Folge p-fach integrierbarer Funktionen $(f_n) \subseteq \mathcal{L}^p(S, \mathbb{R})$ und zwei einzelne Funktionen f, $g : S \to \mathbb{R}$. Dann gelten die folgenden beiden Aussagen:

(i) Ist auch $f \in \mathcal{L}^p(S, \mathbb{R})$ p-fach integrierbar und konvergiert (f_n) μ-fast überall punktweise gegen f (d.h. für $K := \{\, x \in S \mid (f_n(x)) \to f(x) \,\}$ gilt $\mu(S \setminus K) = 0$) dann sind die folgenden beiden Aussagen äquivalent

   (a) Die Folge (f_n) konvergiert *im p-ten Mittel* gegen f, d.h. es gilt

$$\lim_{n \to \infty} \|f - f_n\|_p = \; 0$$

(b) Die p-Normen der f_n konvergieren gegen die p-Norm von f, d.h.

$$\lim_{n\to\infty} \|f_n\|_p = \|f\|_p$$

(ii) *Satz von der majorisierten Konvergenz:* Sei $g \in \mathcal{L}^p(S, \mathbb{R}^+)$ positiv, integrierbar und μ-fast überall eine Majorante der f_n, d.h. es gelte

$$\mu\left(S \setminus \bigcap_{n\in\mathbb{N}} \{\, x \in S \mid |f_n(x)| \leq g(x) \,\}\right) \;=\; 0$$

Weiterhin sei (f_n) μ-fast überall punktweise konvergent, d.h. für $K := \{\, x \in S \mid (f_n(x)) \text{ konvergiert} \,\}$ gelte ebenfalls $\mu(S \setminus K) = 0$. Dann bezeichnen wir die punktweise Grenzfunktion der f_n mit

$$f(x) \;:=\; \begin{cases} \lim_{n\to\infty} f_n(x) & \text{für } x \in K \\ 0 & \text{für } x \notin K \end{cases}$$

Dann ist wiederum $f \in \mathcal{L}^p(S, \mathbb{R})$ p-fach integrierbar und (f_n) konvergiert sogar im p-ten Mittel gegen f, d.h. es gilt auch

$$\lim_{n\to\infty} \|f - f_n\|_p = \; 0$$

Satz 17.13: [Bauer, MundI-Theorie, 17.1 bis 17.10]

(i) Sei $(S, \mathcal{A}, \mu)$ ein Maßraum und $\omega \in \Omega(S)$ eine Belegungsfunktion auf S. Dann erhalten wir ein weiteres Maß $\omega\mu$ auf S, durch:

$$\omega\mu \;:\; \mathcal{A} \to [0, \infty] \;:\; A \mapsto \int_A \omega \, d\mu$$

Und ist $f : S \to \mathbb{R}$ eine messbare Funktion, dann ist f genau dann $(\omega\mu)$-integrierbar, wenn das Produkt $f\omega$ μ-integrierbar ist. In diesem Fall gilt die Gleichheit der Integrale:

$$\int f \, d(\omega\mu) \;=\; \int f\omega \, d\mu$$

(ii) *Satz von Radon-Nikodym:* Ist $(S, \mathcal{A}, \mu)$ ein Maßraum mit σ-endlichem Maß μ und ist $\nu : \mathcal{A} \to [0, \infty]$ ein weiteres Maß auf S, dann sind die filgenden Aussagen äquivalent:

(a) ν besitzt eine Dichte bezüglich μ - d.h. es gibt eine Belegungsfunktion $\omega \in \omega(S)$ so dass $\nu = \omega\mu$ ist.

(b) Jede μ-Nullmenge ist auch eine μ-Nullmenge. D.h. für alle messbaren Mengen $N \in \mathcal{A}$ gilt die Implikation:

$$\mu(N) \;=\; 0 \quad\Longrightarrow\quad \nu(N) \;=\; 0$$

In diesem Fall schreiben wir auch $\nu \leq \mu$. Ist nun ν ein endliches Maß, dann ist dazu weiterhin äquivalent:

(c) ν ist stetig bezüglich μ, d.h. es gibt Zahlen δ und $\varepsilon > 0$, so dass für alle messbaren Mengen $A \in \mathcal{A}$ gilt:

$$\mu(A) \;\leq\; \delta \quad\Longrightarrow\quad \nu(A) \;\leq\; \varepsilon$$

Satz 17.14: [Bauer, MundI-Theorie, 19.1 bis 19.4]

(i) Sei $(S, \mathcal{A}, \mu)$ ein Maßraum und in $(T, \mathcal{B})$ sei T eine nicht-leere Menge und $\mathcal{B}$ eine σ-Algebra auf T. Schließlich sei $\varphi : S \to T$ eine messbare Abbildung. Damit induziert φ das Bildmaß $\varphi_*\mu : \mathcal{B} \to [0, \infty]$ auf T, vermöge $\varphi_*\mu(B) := \mu(\varphi^{-1}(B))$. Ist nun $\omega \in \Omega(T)$ eine Belegungsfunktion auf T, dann gilt

$$I^*_{\varphi_*\mu}(\omega) \;=\; I^*_\mu(\omega\varphi)$$

wobei $\omega\varphi : S \to [0, \infty] : x \mapsto \omega(\varphi(x))$ wieder die Verkettung bezeichnet. Und ist $g : T \to \mathbb{R}$ eine $(\varphi_*\mu)$-integrierbare Funktion, dann ist $g\varphi$ auch μ-integrierbar, d.h. es gilt die Implikation

$$g \in \mathcal{L}_{\varphi_*\mu}(T, \mathbb{R}) \quad\Longrightarrow\quad g\varphi \in \mathcal{L}_\mu(S, \mathbb{R})$$

[Ist φ bijektiv und auch $\varphi^{-1} : T \to S$ messbar, dann gilt auch die umgekehrte Implikationsrichtung]. In diesem Fall besteht dann auch die Gleichheit der Integrale

$$\int_T g \, d(\varphi_*\mu) \;=\; \int_S g\varphi \, d\mu$$

(ii) *Transformationssatz:* Seien U und $V \subseteq \mathbb{R}^d$ zwei offene Teilmengen des $\mathbb{R}^d$ und sei $\Phi : U \to V$ ein C^1-Diffeomorphismus (d.h. Φ ist bijektiv und sowohl Φ als auch Φ^{-1} sind stetig differenzierbar). Wir fixieren das Lebesgue-Maß λ^d auch $\mathbb{R}^d$, dann gilt

$$g \in \mathcal{L}(V, \mathbb{R}) \quad\Longleftrightarrow\quad g\Phi \in \mathcal{L}(U, \mathbb{R})$$

Und in diesem Fall besteht die Gleichheit der folgenden Integrale (dabei bezeichnet $d\Phi$ die *Jacobi-Matrix* von Φ)

$$\int_V g \, d\lambda^d \;=\; \int_U g\Phi \cdot |\det(d\Phi)| \, d\lambda^d$$

Satz 17.15: [Bauer, MundI-Theorie, 23.2 und 23.7]
Sind $(S, \mathcal{A}, \mu)$ und $(T, \mathcal{B}, \nu)$ zwei σ-endliche Maßräume, dann betrachten wir den Produktraum $(S \times T, \mathcal{A} \otimes \mathcal{B}, \mu \otimes \nu)$. Sind nun $s \in S$, $t \in T$ und $P \in \mathcal{A} \otimes \mathcal{B}$ fixiert, dann bezeichnen wir:

$$\begin{aligned} P_s &:= \{\, t \in T \mid (s, t) \in P \,\} \\ P_t &:= \{\, s \in S \mid (s, t) \in P \,\} \end{aligned}$$

Dann sind $P_s \in \mathcal{B}$ und $P_t \in \mathcal{A}$ messbar und damit sind dann die Abbildungen $S \to [0,\infty] : s \mapsto \nu(P_s) \in \Omega(T)$ und $T \to [0,\infty] : t \mapsto \mu(P_t) \in \Omega(S)$ zwei Belegungsfunktionen auf T bzw. auf S.

Satz von Fubini: Ist nun $f \in \mathcal{L}(S \times T, \mathbb{R})$ eine integrierbare Funktion auf dem Produktraum $S \times T$, dann bezeichnen wir weiterhin

$$\begin{aligned} f_s &: T \to \mathbb{R} \;:\; t \mapsto f(s,t) \\ f_t &: S \to \mathbb{R} \;:\; s \mapsto f(s,t) \end{aligned}$$

Dann sind $f_s \in \mathcal{M}(T, \mathbb{R})$ und $f_t \in \mathcal{M}(S, \mathbb{R})$ wiederum messbar und fast überall integrierbar. D.h. bezeichnen wir $L := \{\, s \in S \mid f_s \in \mathcal{L}(T, \mathbb{R}) \,\}$ beziehungsweise $M := \{\, t \in T \mid f_t \in \mathcal{L}(S, \mathbb{R}) \,\}$, dann gilt $\mu(S \setminus L) = 0$ und $\nu(T \setminus M) = 0$. Und damit gilt dann die Gleichheit der Integrale

$$\begin{aligned} \int_{S \times T} f \, d(\mu \otimes \nu) &= \int_L \left(\int_T f(s,t) \, d\nu(t) \right) d\mu(s) \\ &= \int_M \left(\int_S f(s,t) \, d\mu(s) \right) d\nu(t) \end{aligned}$$

Beispiel 17.16:
Die Trias aus Transformationssatz, Satz von Fubini und Fundamentalsatz der Analysis ist in der Lage die meisten Integrale effektiv zu berechnen. Wir wollen hier ein einfaches Beispiel geben, wie diese kombiniert werden können: Dazu starten wir mit einer Funktion $g : D \to \mathbb{R}$ auf der Kreisscheibe vom Radius R in der Ebene $D := \left\{ (x,y) \in \mathbb{R}^2 \mid x^2 + y^2 \leq R^2 \right\}$. Alternativ hätten wir die offene Kreisscheibe $D^\circ = \left\{ (x,y) \in \mathbb{R}^2 \mid x^2 + y^2 < R^2 \right\}$ zu Grunde legen können. Wir können D weitgehend parametrisieren, mit Hilfe von

$$\Phi \;:\; U \to V \;:\; (r, \alpha) \mapsto (r \cdot \cos(\alpha), r \cdot \sin(\alpha))$$

Dabei ist $U :=]0, R[\times]0, 2\pi[$ und $V := D^\circ \setminus \{\, (x,0) \mid x \geq 0 \,\}$. Diese Einschränkung ist notwendig, damit Φ bijektiv und damit ein C^1-Diffeomorphismus wird. Die Jacobi-Matrix von Φ ist dann gegeben, durch

$$d\Phi(r, \alpha) = \begin{pmatrix} \cos(\alpha) & -r \cdot \sin(\alpha) \\ \sin(\alpha) & r \cdot \cos(\alpha) \end{pmatrix}$$

und damit ist $\det(d\Phi) = r$ unabhängig von α. Ist g messbar und fast-überall beschränkt, dann ist g integrierbar (da D ein endliches Maß hat), und da $D \setminus V$ eine Nullmenge ist, gilt die Gleichheit der Integrale

$$\int_D g \, d\lambda^2 = \int_V g \, d\lambda^2$$

Mit Hilfe des Transformationssatzes können wir das Integral über die geschlitzte Kreisscheibe V aber zurück ziehen, auf das offene Rechteck U

$$\int_V g \, d\lambda^2 = \int_U g\Phi \cdot |\det(d\Phi)| \, d\lambda^2$$

$$= \int_0^R \int_0^{2\pi} g\Phi(r,\alpha) \cdot r \, d(r,\alpha)$$

Und da g integrierbar ist erlaubt uns nun der Satz von Fubini das zweidimensionale Integral sukzessive als zwei eindimensionale Integrale auszuwerten. Insgesamt erhalten wir also die Möglichkeit das Integral zu berechnen

$$\int_D g \, d\lambda^2 \quad = \quad \int_0^R r \left(\int_0^{2\pi} g\Phi(r,\alpha) \, d\alpha \right) dr$$

Letzteres ist mit dem Fundamentalsatz der Analysis zumeist gut möglich. Besonders einfach ist der Fall für $g(x,y) = f(x^2+y^2)$ für eine stetige Funktion $f : [0, R^2] \to \mathbb{R}$. Denn in diesem Fall ist g rotationssymmetrisch und es gilt $g\Phi(r,\alpha) = f(r^2)$ ist unabhängig von α. Damit reduziert sich das innere Integral also auf den Faktor 2π. Ist nun F die Stammfunktion von f mit $F(0) = 0$, dann gilt $(F(r^2))' = 2r \cdot F'(r^2) = 2r \cdot f(r^2)$. Und damit können wir das obige Integral also explizit angeben zu

$$\int_D f(x^2+y^2) \, d\lambda^2(x,y) \quad = \quad \pi \int_0^R 2r \cdot f(r^2) dr \quad = \quad \pi F(R^2)$$

Kapitel 18

Open Publication License

I. Copyright
The copyright to each Open Publication is owned by its author(s) or designee.

II. Scope of License
The following license terms apply to all Open Publication works, unless otherwise explicitly stated in the document. Mere aggregation of Open Publication works or a portion of an Open Publication work with other works or programs on the same media shall not cause this license to apply to those other works. The aggregate work shall contain a notice specifying the inclusion of the Open Publication material and appropriate copyright notice.

Severability: If any part of this license is found to be unenforceable in any jurisdiction, the remaining portions of the license remain in force.

No Warranty: Open Publication works are licensed and provided äs is"without warranty of any kind, express or implied, including, but not limited to, the implied warranties of merchantability and fitness for a particular purpose or a warranty of non-infringement.

III. Requirements on both unmodified and modified Versions
Any publication in standard (paper) book form shall require the citation of the original publisher and author. The publisher and author's names shall appear on all outer surfaces of the book. On all outer surfaces of the book the original publisher's name shall be as large as the title of the work and cited as possessive with respect to the title.

IV. Requirements on Modified Works
All modified versions of documents covered by this license, including translations, anthologies, compilations and partial documents, must meet the following requirements:

- The modified version must be labeled as such.
- The person making the modifications must be identified and the modifications dated.
- Acknowledgement of the original author and publisher if applicable must be retained according to normal academic citation practices. The location of the original unmodified document must be identified.
- The original author's (or authors') name(s) may not be used to assert or imply endorsement of the resulting document without the original author's (or authors') permission.

V. Good-Practice Recommendations

In addition to the requirements of this license, it is requested from and strongly recommended of redistributors that:

- If you are distributing Open Publication works on hardcopy or CD-ROM, you provide e-mail notification to the authors of your intent to redistribute at least thirty days before your manuscript or media freeze, to give the authors time to provide updated documents. This notification should describe modifications, if any, made to the document.
- All substantive modifications (including deletions) be either clearly marked up in the document or else described in an attachment to the document. Finally, while it is not mandatory under this license, it is considered good form to offer a free copy of any hardcopy and CD-ROM expression of an Open Publication-licensed work to its author(s).

VI. License Options

Distribution of the work or derivative of the work for commercial purposes is prohibited, unless prior permission is obtained from the copyright holder in written form.

Für Anmerkungen, Hinweise und Korrekturen bin ich immer dankbar.
Sie können mich per Mail kontaktieren, oder besuchen Sie meine Homepage:

abzeidler@gmx.de, bzw.

https://my.cloudme.com/#zeidlerweb

Copyright (C) 6. April 2021

by Andreas Bernhard Zeidler